STATISTICAL POST-PROCESSING OF THE AIR POLLUTION MODEL SIMAIR

HELÉNE ALPFJORD

Master's thesis
2014:E2

LUND UNIVERSITY

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

TYP AV DOKUMENT	DOKUMENTBETECKNING
☒ Examensarbete ☐ Kompendium ☐ Delrapport ☐ Rapport	**LUTFMS—3235-2014**

INSTITUTION

Matematikcentrum. Matematisk statistik, Lunds universitet, Box 118, 221 00 LUND

FÖRFATTARE

Heléne Alpfjord

DOKUMENTTITEL OCH UNDERTITEL

Statistical post-processing of the air pollution model SIMAIR

SAMMANFATTNING

Air pollution is a serious problem today, both from an environmental and from a health point of view. Especially in cities, particles smaller than 10 µm in aerodynamic diameter (PM_{10}) can reach high concentrations. These particles are dangerous, even at low concentrations, since they are small enough to enter the lungs.

In order to estimate the concentration of air pollutants, different measurements and air pollution models can be used. A combination of model data and measurements allows for the assessment of air pollution concentration over larger areas with a lower degree of uncertainty. Statistical post-processing is one approach to combining model data and measurements.

SIMAIR is a Swedish system of models that uses meteorological data, emission data and dispersion models on different geographical scales to calculate the concentration of air pollutants on regional, urban and local levels.

The aim of this Master's Thesis is to study different statistical post-processing methods and to examine their adequacy with regards to dealing with air quality models. One method, Support Vector Regression, is implemented and analysed based on the results from the SIMAIR model.

The compound that is examined is PM_{10}.

The statistical post-processing method is developed based on data from Hornsgatan in Stockholm from the year 2007 to 2009. This method is then validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg.

The results are promising for all three sites; improvements are seen for almost all statistical indicators used to examine model performance.

NYCKELORD

DOKUMENTTITEL OCH UNDERTITEL - SVENSK ÖVERSÄTTNING AV UTLÄNDSK ORIGINALTITEL

UTGIVNINGSDATUM	ANTAL SID	SPRÅK
år 2014 mån 1		☐ svenska ☒ engelska ☐ annat

ÖVRIGA BIBLIOGRAFISKA UPPGIFTER	ISSN
	ISBN
	2014:E2

Abstract

Air pollution is a serious problem today, both from an environmental and from a health point of view. Especially in cities, particles smaller than 10 µm in aerodynamic diameter (PM_{10}) can reach high concentrations. These particles are dangerous, even at low concentrations, since they are small enough to enter the lungs.

In order to estimate the concentration of air pollutants, different measurements and air pollution models can be used. A combination of model data and measurements allows for the assessment of air pollution concentration over larger areas with a lower degree of uncertainty. Statistical post-processing is one approach to combining model data and measurements.

SIMAIR is a Swedish system of models that uses meteorological data, emission data and dispersion models on different geographical scales to calculate the concentration of air pollutants on regional, urban and local levels.

The aim of this Master's Thesis is to study different statistical post-processing methods and to examine their adequacy with regards to dealing with air quality models. One method, Support Vector Regression, is implemented and analysed based on the results from the SIMAIR model.

The compound that is examined is PM_{10}.

The statistical post-processing method is developed based on data from Hornsgatan in Stockholm from the year 2007 to 2009. This method is then validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg.

The results are promising for all three sites; improvements are seen for almost all statistical indicators used to examine model performance.

Keywords

Statistical post-processing, air pollution, Support Vector Regression, machine learning.

Preface

This Master's Thesis was written during the autumn term of 2013 at the Professional Service Department at SMHI (Swedish Meteorological and Hydrological Institute) in Norrköping. I have worked in the methodology and modelling development group, which presently focuses a lot on statistical post-processing for different applications.

There is currently a lot of international cooperation in order to develop a standardized frame work for air quality modelling. Different ways of combining model data and observations are discussed internationally. This thesis is a contribution to that area, where statistical post-processing is applied to the air pollution model SIMAIR.

I want to give my many thanks to my supervisor Mikael Magnusson, PhD in meteorology, who at all times has encouraged me and given me valuable feedback. I also want to thank the rest of my group at SMHI – Fredrik Karlsson, an engineer with remarkable curiosity and helpfulness, Emelie Karlsson, a critically thinking meteorologist whose expertise I have used many times, and Johan Södling, a fellow statistician whom I have received many brilliant pieces of advice from. Thanks to them I have had a wonderful time in Norrköping.

I also want to show my appreciation to the researchers Gunnar Omstedt and Stefan Andersson, two experts on SIMAIR who have gladly answered all my questions and provided me with data.

To my examiner Johan Lindström, PhD in statistics, I owe many thanks for academic input and support in the process of writing my thesis.

Finally I would like to thank my fiancé Marcus Posada for all his loving support and valuable comments.

Populärvetenskaplig sammanfattning

Luftföroreningar skapar många problem idag, framförallt i städerna. Små partiklar, PM_{10}, orsakar folkhälsoproblem som hjärt-, kärl- och lungsjukdomar även vid låga koncentrationer. Huvudsakliga antropogena källor i Sverige är utsläpp från bränsleförbränning och vägslitage, framförallt på grund av dubbdäck.

Vid SMHI används luftmiljömodeller och mätningar för att undersöka luftmiljön. SIMAIR är ett system av modeller som bland annat används för att modellera koncentrationer av olika föroreningar i gaturum. I det här examensarbetet görs ett försök att förbättra SIMAIR genom att använda statistisk efterbearbetning. Olika statistiska metoder undersöks och en, kallad stödvektorsregression (Support Vector Regression, SVR), implementeras och valideras.

SVR är en kraftfull maskininlärningsmetod som är förhållandevis enkel att använda, tack vare tillgång till öppna källkodsprogram till exempel i Python. Enbart ett par parametrar behöver justeras i algoritmen. Metoden kan hantera olinjäriteter och är bra på att undvika överanpassning till träningsdata. För att SVR ska fungera så krävs träningsdata som är representativ för den data som ska predikteras. Ändras förhållanden mellan träningsdata och prediktionsdata – såsom dubbdäcksanvändning eller meteorologiska parametrar – så har inte metoden haft möjlighet att lära av de fallen och presterar därmed sämre.

Tre platser undersöks – Hornsgatan i Stockholm används för att utveckla den statistiska modellen medan E6 vid Gårda i Göteborg och Västra Esplanaden i Umeå används för validering. Gemensamt för dessa tre är att det finns högkvalitativ data från flera års mätningar av $PM_{10.}$

Resultaten är lovande för alla tre platser, med förbättringar för i princip alla statistiska indikatorer som används vid modellutvärdering. Årsmedelvärdet korrigeras väl med SVR, och den linjära korrelationskoefficienten mellan modell och mätning för dygnsmedelvärden ökar till exempel från 0,6 till 0,76 för data från Gårda.

Känsligheten i att använda observerad meteorologisk data jämfört med meteorologisk data i rutnät, interpolerad från modell och observationer, som inparametrar i SVR undersöks. Metoden visar sig vara okänslig för dessa skillnader.

En intressant vidareutveckling av examensarbetet vore att försöka använda närliggande platser som träningsdata för att prediktera på platser där ingen mätning sker. Detta eftersom SIMAIR strävar efter att delvis vara ett komplement till mätningar.

Table of contents

1 Introduction

Air pollution is a serious problem today, both from an environmental and a health point of view. In order to estimate the concentration of important pollutants in different surroundings both measurements and models are used.

There are about 40 observation sites in streetscapes and urban backgrounds throughout Sweden. The sites in Stockholm, Gothenburg and Umeå provide measurements every hour whilst most other sites only measure daily mean concentrations.

There are uncertainties in several components when modelling air pollution. For example there are approximations in the dispersion models, uncertainties in the meteorological data (both due to models and observations) as well as in the emission data, the traffic data and in the physical impact from, for example, streetscapes and land use.

The Forum for Air Quality Modelling (FAIRMODE) was established in 2008 by the European Environment Agency (EEA) and the European Commission Joint Research Centre (JRC) (European Environment Agency 2011). The objective of the Forum is to increase the use of models in air quality assessment, and to examine and share information about modelling tools for policy purposes. Harmonizing the modelling practices in Europe will facilitate cooperation and decision making (Forum for Air Quality Modelling 2013).

SIMAIR is a Swedish system of models that uses meteorological data, emission data and dispersion models on different geographical scales to calculate the concentration of air pollutants from the regional, urban and local levels. The urban concentrations are considered to be mean values above roofs in squares of 1 x 1 km (Andersson and Omstedt 2009, p. 34). To adjust model results so that they can be compared to measurements from arbitrary places, the use of statistical methods is required. The adjustment is motivated specially as European norms of air quality are based on pointwise measurements and the models aim to, partly, provide a substitute for measurements.

1.1 Aim

The aim of this Master's Thesis is to study different statistical post-processing methods and their adequacy with regards to dealing with air pollution dispersion calculations. One method is implemented and analysed. The compounds that are examined are particles smaller than 10 µm in aerodynamic diameter (PM_{10}).

The statistical post-processing method is developed based on data from Hornsgatan in Stockholm during the years 2007 to 2009. This method is then validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg.

Another part of this project is to look at the sensitivity in using measured versus gridded meteorological data as input to a statistical model and to examine uncertainties in different components of the air pollution model.

1.2 Outline of the report

Initially a general description of PM_{10} and the air pollution model SIMAIR is accounted for in Section 2. Then a theory section follows, containing statistical learning, Support Vector Machines, validation and correlation. In Section 4 the implementation of the statistical post-

processing model is described and in Section 5 indicators of model performance are presented. The data for Hornsgatan in Stockholm, which are used to implement the statistical model, are presented thoroughly in Section 6, together with short descriptions of the validation data for Västra Esplanaden in Umeå and Gårda in Gothenburg. In Section 7 the results are presented, both for Hornsgatan and for the validation data. A discussion and conclusions are given in Section 8. Finally some appendices are attached; they contain a theory section on optimization and additional results which are too detailed to be included in the results section.

2 Air pollutant PM_{10}

PM_{10} are particles with an aerodynamic diameter smaller than 10 µm. Particles with aerodynamic diameters smaller than 10 µm are defined as particles with settling speeds less than the speed of a spherical particle, 10 µm in diameter, with a density of $1 g/cm^3$ (Raabe 1976).

A large part of the particles in Sweden comes from emissions of sulphates and nitrates from combustion of biofuels, oil and from wear on roadways, especially due to studded tires. In cities the concentration of PM_{10} can reach very high levels (Swedish Environmental Protection Agency 2011, p. 10). There are natural sources of PM_{10} as well, for example volcanoes, sea spray and windborne dust (Seinfeld and Pandis 2006, p. 55).

There are serious health aspects related to exposure to PM_{10}, even at low concentrations, since the particles are small enough to enter the lungs. Particles smaller than 2.5 µm can, due to their small mass, be transported far and are more dangerous since they enter further into the lungs than larger particles. Cardiovascular and pulmonary diseases increase with exposure and according to Swedish studies 3000-5000 people die prematurely due to particles in Sweden every year (Naturvårdsverket 2013a).

2.1 Clean air – an environmental objective

The Swedish Parliament has chosen 16 environmental objectives that should be met by 2020. One of these objectives is Clean air, which states that "The air must be clean enough not to represent a risk to human health or to animals, plants or cultural assets" (Swedish Environmental Protection Agency 2011, p. 10).

For PM_{10} the proposed maximum yearly and daily mean concentrations are 15 $\mu g/m^3$ and 30 $\mu g/m^3$, respectively. The proposed maximum concentrations of PM_{10} are generally exceeded in streetscapes and sometimes even in urban backgrounds in southern Sweden (Naturvårdsverket 2012).

2.2 Air quality standards

There are current standards of maximum levels of PM_{10} supported by the Swedish Environmental Code. The maximum daily mean concentration of 50 $\mu g/m^3$ is not to be exceeded more than 35 times per year. The maximum yearly mean is 40 $\mu g/m^3$ (Svensk författningssamling 2010, p. 4). Note that these standards are legislated as opposed to the environmental objectives in Section 2.1 above. The municipalities and authorities are mainly responsible for assuring that the standards are met. If they are not met, a program of measures is drafted (Naturvårdsverket 2013b).

2.3 Air pollution models at SMHI

SIMAIR is a system of models using meteorological data, traffic data, emission data and dispersion models for air pollutants on different geographical scales. SIMAIR calculates the concentration of different pollutants at different scales and adds together the regional, urban and local contributions to a single estimate.

SIMAIR is used by municipalities to assess air quality in cities and villages, and compare concentrations of pollutants to current air quality standards (SMHI 2012). SIMAIR plays an important role in testing different scenarios, for example how future concentrations of air pollutants in Swedish cities are affected by different actions (Omstedt et al. 2012). Epidemiological studies concerning health effects of long term particle exposure is another use of SIMAIR (Gidhagen et al. 2013).

The regional concentration contribution in SIMAIR is computed using the regional background dispersion model MATCH, emission data from EMEP (European Monitoring and Evaluation Programme) and meteorological data from HIRLAM. The estimated urban background concentrations are calculated by the urban dispersion model BUM. The local additions are modelled using, for example, road and traffic information, chimney emissions and meteorological data. The components of SIMAIR are shown in Figure 1 below (Andersson and Omstedt 2012, p. 7).

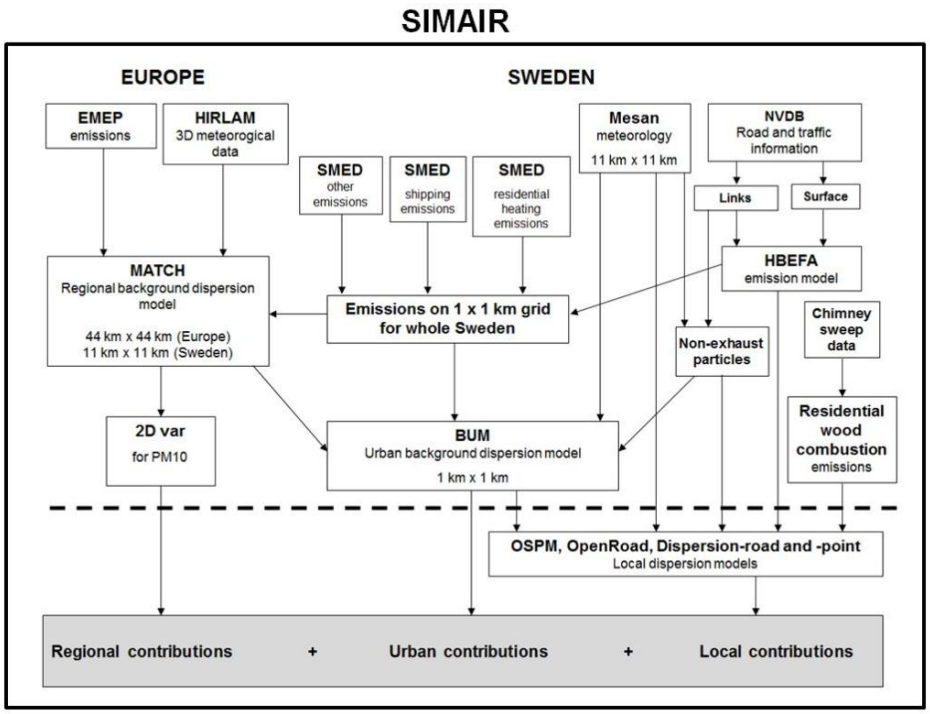

Figure 1. Data and dispersion models used in SIMAIR. Below the dotted line is where the local concentration contributions are calculated.

2.3.1 Regional concentrations

The regional dispersion model MATCH was developed in the 1980's, after the nuclear accident at Tjernobyl, to model the spread of radioactivity (Langner et al. 1998). The model is used in many areas dealing with air pollution, for example problems concerning acidification, eutrophication and low level ozone.

The resolution in the model is dependent of the weather model resolution. For Europe the model is run at a resolution of 44 km x 44 km x 100 m and in Sweden 11 km x 11 km x 100 m. Using discrete time steps (every third minute) the concentration in each gridbox is updated by calculating deposition, chemical transformation and mixing. 3-dimensional meteorological data together with emission data are used as input in MATCH (Robertson, Langner and Engardt 1999). It is an off-line model and there is no feed-back in the relation between air

pollution and weather. For example the effect of high particle concentrations on weather is ignored here.

Validation of the model is difficult, since the gridboxes are very large and the model concentrations are mean values, whilst measurements are pointwise and the concentrations can differ significantly within each gridbox.

EMEP collects emission data on 50 x 50 km resolution. It measures air quality and models the atmospheric transport and deposition of air pollutants. The database is updated every year with a two year delay, e.g. the emission database for 2008 was completed in 2010 (EMEP 2013).

HIRLAM (High Resolution Limited Area Model) is a regionally scaled weather forecasting model. Every six hours predictions are made for two days ahead. The spatial resolution differs depending of area of application, but for the MATCH model used in this application the resolution is 11 x 11 km (Andersson and Omstedt 2012, p. 8). The forecasts in HIRLAM include among others temperature, wind speed and relative humidity. The model was developed jointly by the weather services in Sweden, Norway, Finland, Denmark, Iceland, Ireland, the Netherlands and Spain (Unden et al. 2002).

2.3.2 Urban concentrations

The urban model BUM estimates the concentration of air pollution as the mean concentration in squares of 1 x 1 km above roofs. Emission data originate from SMED (Svenska MiljöEmissionsData), and are updated every year (SMED 2013). SMED is a Swedish database with higher resolution than EMEP. Meteorological data come from MESAN and STRÅNG and are used as input to BUM. In the model calculations BUM distinguishes between emission sources from ground level and from higher levels.

MESAN (meso-scale analysis) is a meteorological analysis model that updates every hour (every third hour when it is used in SIMAIR, and data are then interpolated to hourly values inside the SIMAIR model) and has a spatial resolution of 11 x 11 km. Meteorological observations are made in several places, however as the number of observation stations are limited, MESAN is used in order to combine meteorological models and observations using optimal interpolation techniques. The physical properties of the weather parameters are handled by models similar to those in HIRLAM. No forecasting is done here (Häggmark et al. 2000).

STRÅNG is a mesoscale model for solar radiation (Landelius, Josefsson and Persson 2001) that calculates for example Global Irradiance (GI).

2.3.3 Local concentrations

There are two different models for the local scale dispersion, depending on if a street has buildings on one or both sides (streetscape), or if the surroundings are more open. In streetscapes the dispersion model OSPM (Operational Street Pollution Model) is used, otherwise OpenRoad is chosen.

Local emission factors are collected from HBEFA (Hand Book Emission Factors for Road transport) using traffic information from NVDB (Nationell VägDataBas). Both NVDB and HBEFA are updated every year to adjust for changes in the amount of traffic and the composition of the vehicle fleet. In streets where traffic measurements are done, the traffic

model can be improved by specifying for example the values of yearly daily mean traffic, the proportion of heavy traffic and the use of studded tires.

There is also an emission model for road dust, as well as a model for chimney and local wood heating emissions in SIMAIR (Andersson and Omstedt 2012, p. 8).

3 Theory

In this section statistical learning is described, with focus on Support Vector Machines. First the classification case is introduced, and then Support Vector Regression is presented. Finally there are two subsections containing validation and correlation theory.

3.1 Combined use of models and monitoring data

For a long time monitoring data has been used to examine air quality. An obvious drawback of this approach is that it is difficult to estimate the air quality where there are no measurements. By implementing air quality models one can assess the concentration of air pollution over a much larger area, though the uncertainty can be high. A combination of model data and measurements provides a way of estimating air pollution that combines the benefits from both types of data. This can be done in many ways, and two main categories are data assimilation and data fusion. Several methods can be applied for both categories, such as Kalman filters and residual kriging (Denby and Spangl 2011).

3.1.1 Data assimilation

In data assimilation observation data are used as guidance for the models, by incorporating the observations into the computer model in order to get model states as correct as possible. The physical and chemical characters of the air pollutants described in the dispersion models are retained (Denby and Spangl 2011). Ensemble Kalman filters are examples of data assimilation methods, where an ensemble of model runs is made to estimate the covariance matrix by sample covariance. Another method, used by SMHI, is variational assimilation, for example 2D- and 3D-Var. It can be used to find initial conditions in forecasting by minimizing a cost function and finding a state that fits both model and observations in an optimal way (SMHI 2008). The data assimilation approach is interesting in many ways, however as this thesis treats statistical post-processing, the focus will be on data fusion.

3.1.2 Data fusion

Data fusion methods in air quality modelling are completely statistical, so they do not take physical or chemical laws into account. An advantage is that only the model output is needed, no knowledge regarding the model is required. Examples of methods are regression methods and residual interpolation using either geometric methods such as radial basis functions or geostatistical methods such as kriging (Denby and Spangl 2011).

3.2 Statistical learning

In statistical learning, when predicting an output given inputs, two important cases can be distinguished. If the output is quantitative the prediction is a regression problem and, secondly, if the output is qualitative one has a classification problem. These two types can be considered function approximation tasks and are related (Hastie, Tibshirani and Friedman 2009, p. 10).

The first learning machine (system that can learn from data), the perceptron, was developed in the 1960's by Rosenblatt. It consisted, in its simplest form, of neurons which each has n inputs $x_i, i = 1, \dots, n \in X \subset \mathbb{R}^n$ and one output $y = \{-1, 1\}$. The design of a perceptron is visualized in Figure 2. Each neuron defines two regions of input space X, separated by the hyperplane $y = w \cdot x + b$. During learning the parameters w and b are chosen, however *how* to choose the parameters to achieve generality is the crucial question. In learning theory the connection between training errors and generalization ability is central. If the training errors are too small the model might generalize badly due to overfitting, and if the training errors are too large the model might lose important structure in the data and become too general.

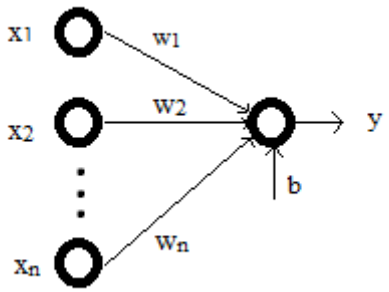

Figure 2. The design of a perceptron.

In the 1980's back-propagation, a technique to find weights of several neurons simultaneously, was developed and enabled training of feed-forward multilayer Neural Networks. The learning theory progressed greatly during the 1970's and 1980's, mostly developed by Vladimir Vapnik and Alexey Chervonenkis. This resulted in a new era of learning machines, including methods focusing on radial basis functions (Vapnik 1999, p. 1-8). The following concepts are central in learning theory.

Let the training data $X \in \mathbb{R}^n$ and the output data $Y \in \mathbb{R}$ be generated independently from a joint distribution $\Pr(X, Y)$. Then a function $f(X)$ is sought after, which predicts Y. In order to penalize errors in training a **loss function** is defined as $L\big(Y, f(X)\big) = (Y - f(X))^2$ (sometimes it is not squared). The expected (squared) prediction error, also called the **risk**, is defined as

$$R(f) = E(L(Y, f(X)) = E(Y - f(X))^2 = \int \big(y - f(x)\big)^2 \Pr(dx, dy)$$

and can be used in order to choose f. By conditioning on X the following expression is obtained:

$$R(f) = E_x E_{Y|X}\big(Y - f(X)\big)^2.$$

The function $R(f)$ is minimized by $f(x) = E(Y|X = x)$, which is called the regression function. The best approximation under squared errors, of Y at $X = x$ is thus the conditional expectation $E(Y|X = x)$ (Hastie, Tibshirani and Friedman 2009, p. 18).

Often the distribution $\Pr(X, Y)$ is unknown, so the risk is approximated by averaging the loss function on the training set, called **empirical risk**:

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)).$$

A problem is that a function f that performs very well on the training data might generalize badly to unseen data. So only minimizing the empirical risk does not imply an optimal performance for test data (unseen data from $\Pr(X, Y)$). Statistical learning theory states that it is necessary to restrict the set of functions from which f is chosen, to a set that has **capacity** appropriate for the amount of training data provided. The VC (Vapnik-Chervonenkis) dimension is a measure of the capacity of function sets, defined as the cardinality of the largest set of points a function set can shatter. The VC theory gives bounds on the test error (the risk), which can be minimized. The minimization, called **structural risk minimization**, depends on empirical risk and the capacity of the function class. More on capacity and structural risk minimization is found for example in Learning with Kernels (Schölkopf and Smola 2002, p. 8-10).

Support Vector Machines (SVMs) are designed to relate to the structural risk minimization when finding an optimal hyperplane for classification and regression. SVMs were originally utilized to classify data into categories, and the powerful method is a form of supervised machine learning. Training data is used to find a hyperplane that separates two classes in an optimal way, see Figure 3. If it is not possible to linearly separate the two classes, the input data is mapped to a higher dimension, called a feature space. The transformation uses kernels, as according to the "kernel trick" any scalar products that need to be calculated can be performed in input space instead of in feature space. This is computationally more efficient. An optimization is done in order to find the best hyperplane, based on minimizing complexity and misclassification. The model can then be used to classify new, unseen data. Commonly used kernels are radial basis functions (RBFs) and polynomials.

The SVMs in their present form were designed by Vladimir Vapnik and Corinna Cortes in 1995. The SVMs have had great success for example in classifying images, handwritten letters and proteins (Christianini and Shawe-Taylor 2012, p. 149-160).

Comparable results can be achieved with Neural Networks, but only if many parameters are optimally tuned by hand. The performance of SVMs depends critically on a few parameters, so qualitative results can be accomplished with less effort, which is a great advantage (Smola and Schölkopf 2004, p. 15).

Support Vector Regression (SVR) uses SVMs for regression analysis. SVR uses transformation into a higher dimensional space in order to find a hyperplane for regression. The optimization minimizes complexity and distance between the hyperplane and training data. Advantages are that SVR is a very general, non-linear, computationally efficient method; it is relatively good at avoiding overfitting as the optimization minimizes model complexity, and the method is easily available since the algorithm is implemented in several open source packages (Marsland 2009, p. 120-129).

3.3 Support Vector Machines

In the following sections Support Vector Machines and Support Vector Regression are described. The basic idea is to find an optimal hyperplane for classification of a linearly separable set. The method is then extended to deal with sets that are not linearly separable, and to solve regression problems.

In order to understand how the optimization problem is solved using the dual problem formulation, a summary of relevant optimization theory is given in Appendix A.

3.3.1 Support Vector Classification

The simplest form of a Support Vector Machine is a maximal margin classifier. It requires linearly separable data in the feature space, and is formulated as a quadratic and convex optimization problem with affine constraints. The problem can thus be solved using the dual formulation, since the duality gap is zero (Appendix A, Proposition 16).

Suppose we have a training set $D = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_1^n$, where y_i indicates which class x_i belongs to. Any hyperplane can be written as $w \cdot x + b = 0$, where w is the normal vector to the plane and $\cdot$ denotes the scalar product. If the data are linearly separable two hyperplanes can be found to separate the two sets, and the optimal hyperplane is located right in between (see Figure 3). The distance between the two planes divided by two is called the functional margin. This margin, a region in which no points are located, can be maximized and the two hyperplanes can be formulated as

$$w \cdot x^+ + b = +1$$
$$w \cdot x^- + b = -1,$$

where x^+ and x^- indicate the points that belong to the positive and negative class, closest to the optimal hyperplane. Here the functional margin (the margin of the function output) is $\frac{+1-(-1)}{2} = 1$ and the geometric margin is

$$\gamma = \frac{1}{2}\left(\frac{w}{||w||_2} \cdot x^+ - \frac{w}{||w||_2} \cdot x^-\right) = \frac{1}{||w||_2}.$$

The following result holds (Christianini and Shawe-Taylor 2012, p. 95):

Proposition 1.

Given a linearly separable training sample

$$S = \left((x_1, y_1), \dots, (x_l, y_l)\right)$$

the hyperplane $f(x) = w \cdot x + b$ *that solves the optimization problem*

$$minimize_{w,b}\ w \cdot w,$$
$$subject\ to\ y_i(w \cdot x_i + b) \geq 1,$$
$$i = 1, \dots, l,$$

realizes the maximal margin hyperplane with geometric margin $\gamma = \dfrac{1}{||w||_2}$.

The inequality constraints above describe the two separating hyperplanes (at equality) and are affine. The optimal, maximal margin hyperplane $f(x)$ solves the optimization problem above and the hyperplane is used to classify points x_i by the decision rule $sgn(f(x_i))$.

Figure 3 (Wikipedia 2008) illustrates a separating hyperplane, together with its margin.

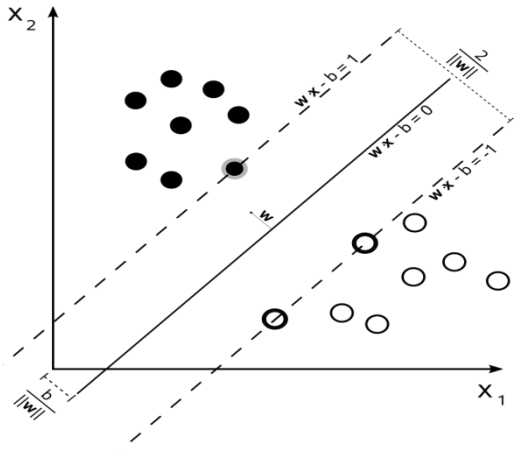

Figure 3. Points of two classes, together with a separating hyperplane and its margin.

3.3.1.1 Dual formulation

To solve the optimization problem the corresponding dual is found. See the section about duality in Appendix A for more theory. Later on (when using kernels for mapping input data which are not linearly separable in input space) solving the dual problem will be simpler than solving the primal. The primal Lagrangian is

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} \alpha_i (y_i (w \cdot x_i + b) - 1), \qquad (1)$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. The dual is found by differentiating $L(w, b, \alpha)$ with respect to b and w and setting the derivatives to 0.

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{l} y_i \alpha_i x_i = 0$$

By substituting $w = \sum_{i=1}^{l} y_i \alpha_i x_i$ and $\sum_{i=1}^{l} y_i \alpha_i = 0$ in the primal (1), $L(w, b, \alpha)$ can be stated as

$$L(w, b, \alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j).$$

The dual problem is formulated below (Christianini and Shawe-Taylor 2012, p. 96).

Proposition 2.

Given a linearly separable training sample

$$S = \big((x_1, y_1), \dots, (x_l, y_l)\big)$$

and suppose α^* *solves the following quadratic optimization problem*:

$$maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j),$$

$$subject\ to \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$\alpha_i \geq 0, i = 1, \dots, l.$$

Then the weight vector $w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i$ *realizes the maximal margin hyperplane with geometric margin*

$$\gamma = \frac{1}{||w^*||_2}.$$

The value of b^* is calculated from the primal constraints as

$$b^* = -\frac{max_{y_{i=-1}}(w^* \cdot x_i) + min_{y_{i=1}}(w^* \cdot x_i)}{2}.$$

According to the Karush-Kuhn-Tucker (KKT) complementary conditions the optimal solution must satisfy

$$\alpha_i^*(y_i(w^* \cdot x_i + b^*) - 1) = 0, \quad i = 1, \dots, l.$$

For the points that are situated exactly on the margin the α_i^* solutions are non-zero. Note that for these points equality hold for the inequality constraints stated in Proposition 1. For all other points the multipliers α_i^* are zero. This implies that only points closest to the margin influence the weight vector. These points are therefore called **support vectors** (Christianini and Shawe-Taylor 2012, p. 94-97).

The solution for the optimal hyperplane f in its dual form is:

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{l} y_i \alpha_i^*(x_i \cdot x) + b^* = \sum_{i \in sv} y_i \alpha_i^*(x_i \cdot x) + b^*,$$

where sv stands for support vectors. The KKT condition also results in:

$$w^* \cdot w^* = \sum_{i \in sv} \alpha_i^*.$$

This implies that the optimal hyperplane has the following geometric margin in its dual formulation:

$$\gamma = \frac{1}{||w||_2} = \left(\sum_{i \in sv} \alpha_i^* \right)^{-\frac{1}{2}}.$$

The fact that the input x_i in its dual form is inside an inner product both in its objective function $W(\alpha)$ and in its decision function $f(x, \alpha^*, b^*)$ allows for the use of kernels in the transformation of data from input space to a feature space in order to find an appropriate hyperplane when data are not linearly separable in input space (Christianini and Shawe-Taylor 2012, p. 97-98).

3.3.1.2 Transformation using kernels

The idea is to transform data that are not linearly separable in input space to a feature space in order to achieve a linear separation. This is done by mapping the training samples x_i by a map $\Phi: X \rightarrow \mathcal{F}$ into a feature space $\mathcal{F}$. The trick is that the scalar product $\Phi(x) \cdot \Phi(z)$ never needs to be computed explicitly if Φ is a kernel. Instead a kernel matrix $K(x,z) = \Phi(x) \cdot \Phi(z)$ is found, allowing each dot product to be replaced by a kernel function. The dot product is thus computed in input space instead of computing dot products for the extended basis vectors (Smola and Schölkopf 2004, p. 3).

The kernel trick is the fundament of Support Vector Machines, without it the calculations would become computationally too expensive. According to Mercer's theorem any symmetric, positive definite function can be used as a kernel. The dot products of the extended basis vectors are substituted by a kernel matrix (a Gram matrix) K.

Common types of kernels are seen in Table 1. It can be noted that sigmoid functions are conditionally positive definite in certain parameters, and are valid kernels for those parameters (Lin and Lin 2003, p. 1).

Table 1. Commonly used kernels in Support Vector Machines.

Polynomials (inhomogeneous or homogeneous)	$K(x,z) = (1 + x \cdot z)^s$ or $K(x,z) = (x \cdot z)^s$
Sigmoid functions	$K(x,z) = \tanh(\kappa x \cdot z - \delta)$
Radial basis functions	$K(x,z) = \exp(-\frac{(x-z)^2}{2\sigma^2})$

There is no objective way of choosing which kernel to use; the present technique is trial and error (Marsland 2009, p. 126-127).

Example: A simple kernel:

Consider the map $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ with $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$. The dot product can be calculated based on the dot product in input space: $\Phi(x_1, x_2) \cdot \Phi(x_1', x_2') = (x \cdot x')^2$. Thus, it suffices to know $K(x_1, x_2) = \Phi(x_1, x_2) \cdot \Phi(x_1', x_2')$, and we do not need to know $\Phi(x)$ explicitly (Smola and Schölkopf 2004, p. 3).

The use of kernels in the maximal margin optimization problem is visualized in Proposition 3 (Christianini and Shawe-Taylor 2012, p. 98-99).

Proposition 3.

Consider a training sample

$$S = \big((x_1, y_1), \dots, (x_l, y_l)\big)$$

that is linearly separable in the feature space implicitely defined by the kernel $K(x,z)$ *and suppose the parameters* α^* *and* b^* *solve the following quadratic optimization*

$problem$:

$$maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j),$$

$$subject\ to \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$\alpha_i \geq 0, i = 1, \dots, l.$$

$Then\ the\ decision\ rule\ given\ by\ sgn\big(f(x)\big), where\ f(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^*,$

$is\ equivalent\ to\ the\ maximal\ margin\ hyperplane\ in\ the\ feature\ space\ implicitely\ defined\ by\ the\ kernel\ K(x, z)\ and\ that\ hyperplane\ has\ geometric\ margin$

$$\gamma = \frac{1}{||w||_2} = \left(\sum_{i \in sv} \alpha_i^* \right)^{-\frac{1}{2}}.$$

3.3.1.3 Soft margin optimization

Often linear separation is not possible, even after a transformation using kernels is performed, due to noisy data. This is handled by a soft margin optimization, where slack variables ξ_i, allow for some misclassified points. The optimization problem can be formulated as a 2-Norm Soft Margin:

$$minimize_{\xi, w, b}\ w \cdot w + C \sum_{i=1}^{l} \xi_i^2$$

$$subject\ to\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, \qquad i = 1, \dots, l,$$

where the parameter C is determined by (cross-)validation. When C is varied the norm $||w||_2$ also varies, so ξ_i is minimized based on the choice of C. The parameter C balances misclassification errors against complexity.

The primal Lagrangian, connected to the 2-Norm Soft Margin problem, is

$$L(w, b, \xi, \alpha) = \frac{1}{2} w \cdot w + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 - \sum_{i=1}^{l} \alpha_i (y_i (w \cdot x_i + b) - 1 + \xi_i),$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. In order to find the dual, the Lagrangian $L(w, b, \xi, \alpha)$ is differentiated and the derivatives are set to 0:

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial w} = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0$$

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial \xi} = C\xi - \alpha = 0$$

$$\frac{\partial L(w,b,\xi,\alpha)}{\partial b} = \sum_{i=1}^{l} y_i\alpha_i = 0,$$

and by resubstitution, $L(w,b,\xi,\alpha)$ is formulated as:

$$L(w,b,\xi,\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \frac{1}{2C}(\alpha \cdot \alpha).$$

The dual problem (with a general kernel) is stated below (Christianini and Shawe-Taylor 2012, p.106):

Proposition 4.

Consider classifying a training sample

$$S = \big((x_1, y_1), \dots, (x_l, y_l)\big),$$

using the feature space implicitely defined by the kernel $K(x,z)$, and suppose the parameters α^* solve the following quadratic optimization problem:

$$maximize\ W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \left(K(x_i, x_j) + \frac{1}{C}\delta_{ij}\right),$$

$$subject\ to \sum_{i=1}^{l} y_i\alpha_i = 0,$$

$$\alpha_i \geq 0, \qquad i = 1, \dots, l.$$

Let $f(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^*$, where b^* is chosen so that $y_i f(x_i) = 1 - \frac{\alpha_i^*}{C}$ for any i with $\alpha_i^* \neq 0$.

Then the decision rule given by $sgn\big(f(x)\big)$ is equivalent to the hyperplane in the feature space implicitely defined by the kernel $K(x,z)$ which solves the optimization problem, where the slack variables are defined relative to the geometric margin

$$\gamma = \left(\sum_{i \in sv}\alpha_i^* - \frac{1}{C}(\alpha^* \cdot \alpha^*)\right)^{-\frac{1}{2}}.$$

This follows from the KKT condition $\alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) = 0, i = 1, \dots, l$, and the relation $C\xi_i = \alpha_i$. This problem is closely related to the optimization problem for the maximal margin; it can actually be seen as a change of kernel $K'(x,z) = K(x,z) + \frac{1}{C}\delta_x(z)$.

There is a 1-Norm Soft Margin as well. The corresponding primal Lagrangian is expressed as:

$$L(w,b,\xi,\alpha,r) = \frac{1}{2}w \cdot w + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{l} r_i\xi_i,$$

where $\alpha_i \geq 0$ and $r_i \geq 0$. The same procedure as for the 2-Norm Soft Margin is done in order to find the corresponding dual, stated in Proposition 5 (Christianini and Shawe-Taylor 2012, p.108).

Proposition 5.

Consider a training sample

$$S = \big((x_1, y_1), \dots, (x_l, y_l)\big)$$

that is linearly separable in the feature space implicitely defined by the kernel $K(x, z)$ *and suppose the parameters* α^* *and* b^* *solve the following quadratic optimization problem*:

$$maximize\ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j),$$

$$subject\ to \sum_{i=1}^{l} y_i \alpha_i = 0\,, \qquad C \geq \alpha_i \geq 0, i = 1, \dots, l.$$

Let $f(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^*$, *where* b^* *is chosen so that* $y_i f(x_i) = 1$ *for any* i *with* $C > \alpha_i^* > 0$.

Then the decision rule given by $sgn\big(f(x)\big)$ *is equivalent to the hyperplane in the feature space implicitely defined by the kernel* $K(x, z)$ *which solves the optimization problem, where the slack variables are defined relative to the geometric margin*

$$\gamma = \left(\sum_{i,j \in sv} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \right)^{-\frac{1}{2}}.$$

The KKT conditions state that if $C > \alpha_i^* > 0$ then $\xi_i^* = 0$ and $y_i(w^* \cdot x_i + b^*) - 1 + \xi_i^* = 0$, which is used above. The points with $\xi_i^* > 0$ are margin errors as their geometric margin is less than $\frac{1}{||w||_2}$, and they can only occur when $\alpha_i^* = C$. These results correspond exactly to the Maximal Margin problem, with the difference that all α_i are bounded from above by C. This constrains the influence of outliers, as they otherwise would have large Lagrange multipliers. The 1-Norm Soft Margin thus has the name Box Constraint.

Both the 1- and 2-Norm Soft Margins are used today. Which margin that gives the best result depends on the data, and the type and amount of noise (Christianini and Shawe-Taylor 2012, p.105-110).

3.3.2 Support Vector Regression

The idea of Support Vector Regression is to find a function $f(x) = w \cdot x + b$ that has at most deviation ε from the actual targets y_i, for all training data x_i. If not possible linearly, the input data can be transformed into a higher dimension to find a hyperplane that has at most ε deviation from each transformed point. The optimal hyperplane is found by minimizing the flatness (norm of w) of the hyperplane and the deviations larger than ε.

The simplest case is the linear problem, formulated as follows:

$$minimize_{w,b}\ \frac{1}{2} w \cdot w$$

$$subject\ to\ \ |y_i - w \cdot x_i - b| \ \leq\ \varepsilon,$$

where input data is on the form $S = \big((x_1, y_1), \dots, (x_l, y_l)\big) \subset \mathrm{X} \times \mathbb{R}$, where for example $\mathrm{X} = \mathbb{R}^d$. Often there is no function $f(x) = w \cdot x + b$ that approximates (x_i, y_i) with a maximum deviation of ε. Similar to the 1-Norm Soft Margin in the classification case, slack variables can be introduced, allowing a few points to deviate more than ε from the hyperplane. The problem is then formulated as:

$$minimize_{\xi,w,b}\ \frac{1}{2} w \cdot w + \ C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$subject\ to\ \ \ y_i - w \cdot x_i - b \ \leq\ \varepsilon + \xi_i,$$

$$w \cdot x_i + b - \ y_i \ \leq\ \varepsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq 0.$$

The parameter C decides the compromise between flatness of $f(x)$ and the penalty of deviations larger than ε. This penalty is called the ε-insensitive loss function $|\xi|_\varepsilon$ and can be stated as:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if} |\xi| \leq \ \varepsilon \\ |\xi| - \ \varepsilon & \text{otherwise.} \end{cases}$$

The same method as for Support Vector Classification is used – state the primal Lagrangian, derive and set derivatives to zero, resubstitute and state the dual. The primal Lagrangian is, in this case:

$$L(w, b, \xi, \eta, \alpha) = \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*)$$

$$- \sum_{i=1}^{l} \alpha_i (\varepsilon + \xi_i - y_i + w \cdot x_i + b)$$

$$- \sum_{i=1}^{l} \alpha_i^* (\varepsilon + \xi_i^* + y_i - w \cdot x_i - b),$$

with constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$.

The derivatives, imposing a stationary point, are:

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial b} = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0$$

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial w} = w - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial \xi_i^{(*)}} = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0,$$

where $\alpha_i^{(*)}$denotes α_i and α_i^*.

The resulting dual problem can be formulated as:

$$maximize\ W(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i \cdot x_j)$$

$$-\varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$subject\ to\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \qquad \alpha_i^{(*)} \in [0, C].$$

The function $f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\,(x_i \cdot x) + b$ describes the hyperplane, but b needs to be computed. The KKT conditions:

$$\alpha_i(\varepsilon + \xi_i - y_i + w \cdot x_i + b) = 0 \qquad (1)$$

$$\alpha_i^*(\varepsilon + \xi_i^* + y_i - w \cdot x_i - b) = 0 \qquad (2)$$

$$(C - \alpha_i)\xi_i = 0 \qquad (3)$$

$$(C - \alpha_i^*)\xi_i^* = 0 \qquad (4)$$

give useful information. Only points with a deviation larger than ε have $\alpha_i^{(*)} = C$. Also $\alpha_i\alpha_i^* = 0$, so if $\alpha_i < C$ then $\xi_i = 0$ and $y_i - w \cdot x_i - b < \varepsilon$. One way of finding b is through the following inequalities:

$\max\{-\varepsilon + y_i - w \cdot x_i | \alpha_i < C \text{ or } \alpha_i^* > 0\} \leq b \leq \min\{-\varepsilon + y_i - w \cdot x_i | \alpha_i > 0 \text{ or } \alpha_i^* < C\}$.

If any $\alpha_i^{(*)} \epsilon (0, C)$, the inequalities become equalities.

For $|f(x_i) - y_i| < \varepsilon$ all $\alpha_i^{(*)}$ have to be zeros in order to fulfil the first two KKT conditions. Therefore only a part of the inputs x_i are used in order to describe the hyperplane, and these are called the support vectors. The support vectors are the points that have a deviation larger than ε from the hyperplane (Smola and Schölkopf 2004, p. 1-3).

Support Vector Regression can be generalized to handle nonlinearities by using kernels, just like for classification. The dual problem can be reformulated as follows (Christianini and Shawe-Taylor 2012, p. 117-118):

Proposition 6.

Suppose a regression is to be performed on a training sample

$$S = \big((x_1, y_1), \ldots, (x_l, y_l)\big)$$

using the feature space implicitely defined by the kernel $K(x, z)$

and suppose the parameters $\widehat{\alpha^*}, \hat{\alpha}$ *and* $\hat{b}$ *solve the following quadratic optimization problem*:

$$maximize\ W(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)\big(\alpha_j - \alpha_j^*\big)K\big(x_i, x_j\big)$$

$$-\varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$subject\ to\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \qquad \alpha_i^{(*)} \in [0, C].$$

$$Let\ f(x) = \sum_{i=1}^{l} (\widehat{\alpha_i} - \widehat{\alpha_i^*})\, K(x_i, x) + \hat{b}, where\ \widehat{b}\ is\ chosen\ so\ that\ f(x_i) - y_i = -\varepsilon,$$

$$for\ any\ i\ with\ 0 < \alpha_i^{(*)} < C.\ Then\ the\ function\ f(x)\ is\ equivalent\ to\ the\ hyperplane$$

$$in\ the\ feature\ space\ implicitely\ defined\ by\ the\ kernel\ K(x,z)\ which\ solves$$

$$the\ optimization\ problem\ above.$$

It can be noted that $f(x)$ contains a kernel function, however w is no longer known explicitly. The parameter $\hat{b}$ is found using the KKT conditions (1) and (2) stated above.

When performing a Support Vector Regression on new input data, the points are projected to the hyperplane determined in training.

3.3.3 Validation

In order to evaluate a learning algorithm a common statistical method divides the known data set into a training set and a validation set, and test the performance gained from the training set on the validation set. This method is called validation, and reduces the problems with overfitting when predicting on an unknown set.

The data set is often divided into three sets: a **training set** for the learning algorithm, a **validation set** on which the algorithm is tested in order to avoid overfitting, and finally a **test set**, which is a set on which prediction is to be made.

There are different types of validation. For example the k-fold cross validation divides the data set into k equally large subsets and rotates so that each subset is used for validation while the other $k - 1$ subsets are used for training. For example the average of the k results can be used for comparison. Drawbacks are that training data and validation data overlap and are dependent, and for large datasets this method can become very computationally expensive.

The hold-out validation is another type of validation. Here the data set is divided into a training set and a validation set, which do not overlap. The disadvantage is that the data in the validation set is never used for training, so important information may be lost, which could lead to higher variance. When having a small data set this type of validation is probably unwise (Refaeilzadeh, Tang and Liu 2008).

For all choices of kernel the parameter C has to be set in Support Vector Machines. This parameter decides the trade-off between the complexity of the hyperplane and the penalty of samples which are far from the hyperplane. A smaller C results in a flatter hyperplane and a higher C aims at finding a hyperplane that fits all samples well.

When using Radial Basis Functions (RBFs) the parameter γ also needs to be set. The kernel is formulated as

$$K(x, z) = \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) = \exp(-\gamma(x - z)^2),$$

so $\gamma = \frac{1}{2\sigma}$, where σ denotes the standard deviation. The parameter γ decides how much influence each training example has. A large γ results in a low standard deviation so only examples very close to the training point are influenced, and vice versa.

Selection of these parameters is not trivial. The most common way is to do a grid search with the selected parameters evaluated using a validation set (Marsland 2009, p. 127). In Section 4.3 the implemented grid search on a log scale is described. A hold-out validation is done in

order to find the optimal combination. The error tolerance ε is another parameter in the SVR that can be tuned with validation.

3.4 Correlation – Pearson's r and Spearman's rank

There are many ways to measure dependence between two data sets X and Y. Pearson's product-moment correlation coefficient, or Pearson's r, is a linear correlation measure calculated as:

$$r_{X,Y} = \frac{1}{n-1} \frac{(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})}}.$$

The correlation coefficient is in the interval [-1, 1], where 1 implies complete positive linear dependence and -1 complete negative linear dependence.

Spearman's rank correlation coefficient also measures dependence between two data sets, based on how well their dependence can be described by a monotone function. A monotone function is either a non-decreasing or a non-increasing function. This correlation coefficient can thus measure non-linear dependencies following a monotone function. In other words it measures if the ordering is the same, e.g. if the fifth element is the largest in both datasets. Spearman's rank is defined as the Pearson's correlation coefficient between the ranked variables in the data sets. When there are no duplicates within each set, Spearman's rank can be computed as:

$$r_{X,Y} = 1 - \frac{6\sum_{i=1}^{n}\left(R_X(i) - R_Y(i)\right)^2}{(n^2-1)n},$$

where n is the number of elements in each set and $R_X(i)$ and $R_Y(i)$ are the ranks for x_i and y_i. The correlation coefficient is in the interval [-1, 1] where 1 denotes complete positive monotone dependence and -1 complete negative monotone dependence (Blom et al. 2005, p. 232-235).

4 Implementation

This section describes the implementation of the Support Vector Regression. The algorithms used in this thesis are implemented in an open source library called Scikit learn.

One part of the model design is to choose relevant explanatory variables as input to the SVR. In Section 6.1.1 the characteristics of Hornsgatan data are described. Based on this analysis the explanatory variables used in the SVR are chosen. Another part of the model design is to select values of the C and γ parameters. This selection, together with how data are adjusted to the algorithm, is described below.

4.1 Scikit learn – Machine learning in Python

Scikit learn is an open source Python module that includes several methods of machine learning, both supervised and unsupervised. The input data are structured as Numpy arrays, where Numpy is an open source library that extends the Python programming language to easily deal with arrays, matrices and mathematical operations (Pedregosa et al. 2011).

The development of Scikit learn is funded mostly by INRIA – the French Institute for Research in Computer Science and Control and by Google. It started in 2007 as a Google Summer of Code project by David Cournapeau and in 2010 INRIA started to lead the project and released the first public version (Scikit learn 2013a).

Scikit learn provides implementations of Support Vector Machine methods, including Support Vector Regression. Scikit learn uses the C library libSVM. One can specify which kernel to use out of linear, polynomial, RBF and sigmoid kernels. If desired one can also define own kernels. The module solves a 1-Norm Soft Margin optimization problem (Scikit learn 2013b).

4.2 Exponential moving average filter

An exponential moving average (EMA) filter is used on the precipitation data in order to obtain a better idea of when streets are wet. After discussions with meteorologists at SMHI, the number of lags was chosen to be 18 hours. This implies that the filtered rain $\bar{x}_k$, which indicates wetness of the street, is influenced by precipitation during the past 18 hours. An EMA filter places more emphasis on recent rainfall by discounting older values exponentially. This can be expressed mathematically:

$$\begin{aligned}\bar{x}_k &= \theta\bar{x}_{k-1} + (1-\theta)x_k = \theta[\theta\bar{x}_{k-2} + (1-\theta)x_{k-1}] + (1-\theta)x_k \\ &= \theta^2\bar{x}_{k-2} + \theta(1-\theta)x_{k-1} + (1-\theta)x_k,\end{aligned}$$

and this can be further expanded by expanding $\bar{x}$ terms. The parameter $\theta \in (0,1)$ is calculated as $\theta = \frac{lags}{lags+1}$. A larger θ gives a larger degree of filtering and a smaller θ expresses less filter influence (Tham 2009).

4.3 Grid search and hold-out validation

In this thesis a Radial Basis Function kernel is used, since it is one of the most common kernel choices. This selection implies that both the values of C (required for all kernels) and γ need to be set. As described in section 3.3.3 a grid search is done together with a hold-out

validation. Originally the grid search was done on the interval $C = \{2^{-5}, 2^{-3}, \dots, 2^{15}\}$ and $\gamma = \{2^{-15}, 2^{-13}, \dots, 2^{5}\}$, but after some testing the intervals can be decreased, as one intuitionally knows which parameter values that are reasonable for the problem in question. The choice of error tolerance ε is not sensitive in this implementation. The default value of $\varepsilon = 0.1$ is used as maximum allowed deviation.

When doing Support Vector Regression in this thesis, three years of data are available – year 2007 to 2009. In order to prepare for the validation, two evaluation years, 2007 and 2008, are divided into a training set and a validation set. The validation set is constructed by randomly picking 25% of the two years, with the remaining 75 % becoming the training set. When doing grid search for C and γ each combination of the two parameters is used for prediction on the validation set. The parameters that result in the smallest MAE for the validation set are then used, and training is redone on the entire evaluation set. The final prediction is made on the test set, the data from 2009.

4.4 Scaling of input vectors

As Support Vector Machines are not scale invariant it is a good idea to scale the input vectors. The scaling performed on the vectors (those that are not indicators) is the following: calculate the mean and the standard deviation of the vectors in the training set, then scale both the training set and the validation set by subtracting the computed mean and dividing by the standard deviation. For the test set, scaling is based on the mean and standard deviation of the entire evaluation set (the training set and validation set put together).

4.5 Three regressions

Data are divided into three intervals based on the model values from SIMAIR in streetscapes. The intervals are decided by dividing the data into large enough parts and by testing different intervals. The final choice fell on the intervals < 25, 25-50 and >50 µg/m^3. Thereafter three regressions are executed, one in each interval. The advantage of implementing several regressions is that each regression will fit the data in its interval better than one regression covering all the concentration values would.

5 Model performance

Air pollution models are increasingly used for policy support, and there is a need for standard methods to evaluate the models. One way to evaluate the quality is to compare model results with measurements; this is called operational model evaluation or statistical performance analysis. Several statistical performance indicators should be used, for example indicators that describe the bias, the correlation, the standard deviation and the root mean square error (RMSE), as they provide information about general model performance. In order to have a quality evaluation of the models for policy use, Model Performance Criteria (MPC) should be defined and fixed. This work is currently in progress (Thunis, Pederzoli and Pernigotti 2012).

5.1 Model Performance Criteria

The European Air Quality Directive, adopted by the European Commission, defines the uncertainty of the models as the largest deviation between calculated and measured concentration for 90 % of the observations, closest to the limit value, over the period in question. This is computed by sorting both the model values and the measurements in increasing order, remove the top 10 %, and compare the measured value closest to the limit value to the model value with the same index. No consideration is thus given the temporal order of the data.

Currently there is only a model quality objective for the yearly mean value of PM_{10}, which is an uncertainty of maximum 50 % for the modelled concentration. The objective for daily concentration has not been decided yet (EU 2008, p. 14). Within FAIRMODE (see Section 1) two statistical indicators were developed to describe model performance; the Relative Percentile Error (RPE) and Relative Directive Error (RDE).

There are several statistical performance indicators used to evaluate air pollution models. Some of the most important are listed below.

- Mean concentration over a certain time period: $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$
- Standard deviation: $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C})^2}$
- Root Mean Square Error: $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)^2}$
- Correlation coefficient: $r = \frac{\sum_{i=1}^{n}(M_i-\bar{M})(O_i-\bar{O})}{\sqrt{\sum_{i=1}^{n}(M_i-\bar{M})^2(O_i-\bar{O})^2}}$
- Relative Percentile Error: $\mathrm{RPE} = \left|\frac{O_p-M_p}{O_p}\right|$
- Relative Directive Error: $\mathrm{RDE} = \left|\frac{O_{LV}-M_{LV}}{LV}\right|$

Here C_i is any concentration, O_i is observed concentration, M_i is modelled concentration, O_p is observed concentration for the percentile in question and M_p is modelled concentration for the same percentile. For example the 98 percentile is the concentration which 98 % of the values fall below and 2 % exceed. O_{LV} is the observed concentration and M_{LV} is the modelled concentration closest to the limit value LV.

Which of the indicators RPE and RDE that is used for evaluation of model quality depends on whether the concentrations are close to the maximum limit. For low hourly and daily mean concentrations the RPE is preferred, but the opposite applies for yearly mean values. So for

high yearly mean concentrations the use of RPE is preferred and for low yearly mean values RDE is better (Andersson S. and Omstedt G. 2012, p. 13-15).

5.1.1 Model performance criteria based on observational uncertainty

FAIRMODE suggests using model performance criteria based on observational uncertainty. The observational uncertainty is defined as $U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(U_r(O_i) \cdot O_i)^2}$,

where U_r is the relative uncertainty for a given concentration and pollutant. In the article by Thunis, Pederzoli and Pernigotti (2012) U_r is assumed to be independent of concentration and is set to 25 % for PM_{10}, according to the data quality objective in the European Air Quality Directive.

Statistical indicators suggested by FAIRMODE are:

- *Normalized RMSE:* $RMSE_U = \frac{RMSE}{2U}$
- *Centred RMSE: CRMSE =* $\sqrt{\frac{1}{N}\sum_{i=1}^{N}((O_i - \bar{O}) - (M_i - \bar{M}))^2}$
- *Normalized Mean Bias:* $NMB = \frac{\bar{M}-\bar{O}}{\bar{O}}$
- *Normalized Mean Standard Deviation:* $NMSD = \frac{\sigma_M - \sigma_O}{\sigma_O}$,

and recommended Model Performance Criteria are:

- $RMSE_U < 1$
- $|NMB| < \frac{2U}{\bar{O}}$
- $|NMSD| < \frac{2U}{\sigma_O}$.

These criteria can be reformulated as criteria for bias, correlation and standard deviation in the following manner:

- for bias: $\frac{\bar{M}-\bar{O}}{2U} < 1$
- for correlation: $\frac{(1-r_{M,O})}{2}(\frac{\sigma_O}{U})^2 < 1$,
- and for standard deviation: $\frac{|\sigma_M - \sigma_O|}{2U} < 1$.

In order to visualize the model performance a Target diagram is used (Thunis, Pederzoli and Pernigotti 2012).

The Target is defined as:

$$Target = \frac{1}{2}\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}U_i^2}},$$

where U_i is the observational uncertainty at each point (Andersson S. and Omstedt G. 2012, p. 13). In the Target diagram the y-axis is $\frac{\text{Bias}}{2U}$ and the x-axis is the centred $RMSE_U$. Because the centred RMSE (CRMSE) is always positive the negative x-axis can be used to provide more information in the plot. The ratio of two CRMSEs, one obtained by assuming perfect correlation ($r_{M,O} = 1$) and the other assuming perfect standard deviation ($\sigma_M = \sigma_O$), is calculated and used to decide which side of the Target diagram the point will be placed on:

$$\frac{CRMSE(r_{M,O}=1)}{CRMSE(\sigma_M=\sigma_O)}=\frac{NMSD}{\sqrt{2(1-r_{M,O})}}\begin{cases}>1 \rightarrow SD\ dominates \rightarrow right\\ <1 \rightarrow r\ dominates \rightarrow left.\end{cases}$$

The relation $CRMSE^2=\sigma_O^2+\sigma_M^2-2\sigma_O\sigma_M r_{M,O}$ is used above (Thunis, Pederzoli and Pernigotti 2012).

The closer the points in the Target diagram are to the origin, the better the model result is. If the points are within a radius of 0.5 ($Target<0.5$) the RMSE is smaller than the uncertainties of the measurement, so no improvement is possible. If $0.5<Target<1$, then RMSE is larger than the observational uncertainty, but the model may still be better than the observations since observational and model uncertainty intervals overlap. If the points are further away than 1 from the origin, then the model describes the concentrations worse than the measurements do, according to the assumption of an observational uncertainty of 25 % (Andersson S. and Omstedt G. 2012, p. 16).

5.2 Delta tool

The DELTA tool is software developed by FAIRMODE with the purpose of evaluating air pollution models according to the standards in the European Air Quality Directive. The DELTA tool uses paired data – modelled and measured data for the same place and time period – and does model diagnostics based on certain model performance criteria. The results are presented in different plots, tables and diagrams in order to get a good overview of the model performance (European Commission JRC 2013).

6 Data

The data used to develop statistical post-processing methods in this study consist of three years of observations and model values from Hornsgatan in Stockholm. In order to evaluate how general the derived model is, data from Västra Esplanaden in Umeå and Gårda in Gothenburg are used for validation. The characteristics of the data from Hornsgatan are thus examined thoroughly in the report, while the data sets from Umeå and Gothenburg are only used for validation.

6.1 Stockholm

There are two observation sites at Hornsgatan in Stockholm, both under the care of the Environmental Department in Stockholm. The sites are used for validation of SIMAIR in streetscapes, and measure concentrations of different pollutants, including PM_{10}. The site used in this thesis is situated on the north side of the street, about three meters above ground, and it measures hourly concentrations. The street is 24 meters wide and the buildings on each side are about 24 meters high. There is heavy traffic on this street, with a yearly daily average of 28 000 vehicles of which about 3 % are heavy vehicles. The maximum velocity allowed is 50 km/h and salt is used when lanes are slippery. The percentage of vehicles with studded tires was 73 % in 2007, 69 % in 2008 and 68 % in 2009 (Andersson and Omstedt 2012, p. 4). The time series used stretch from 2007 to 2009. A ban on studded tires was introduced after 2009, and does not affect the examined years.

There is another observation site at Torkel Knutssonsgatan, situated on a roof 20 meters above ground. This site is 100 meters away from Hornsgatan and measures the urban background concentration of pollutants. This observation site also measures temperature, wind speed, wind direction and humidity.

In Figure 4 the observation site at Hornsgatan is marked with a blue dot, and the station at Torkel Knutssongatan is marked with a red dot.

Figure 4. The observation site at Hornsgatan (streetscape) is marked in blue and the station at Torkel Knutssonsgatan (urban background) in red. Source: OpenStreetMap.

There are meteorological parameters from MESAN available (gridded in 11x11 km squares). Table 2 shows available parameters from the observation site at Torkel Knutssonsgatan and Observatoriekullen, and from the two models MESAN and STRÅNG. There are more meteorological parameters than GI (global irradiance) available for Observatoriekullen, but these are not used in this study.

Table 2. Meteorological parameters used as observations and from MESAN and STRÅNG.

MESAN	Temperature (K)	Wind speed (m/s)	Wind direction (degrees 0-360)	Precipitation (mm)	Humidity (%)	
STRÅNG						GI (W/m^2)
Torkel Knutssonsgatan	Temperature (K)	Wind speed (m/s)	Wind direction (degrees 0-360)	Precipitation (mm)	Humidity (%)	
Observatoriekullen						GI (W/m^2)

There is modelled traffic data for Hornsgatan, based on measurements from SLB (Stockholm och Uppsala Luftvårdsförbund); hourly data on number of vehicles, including number of light and heavy vehicles, are also available. These values are based on a traffic model for which input parameters consist of yearly daily mean traffic, proportion of heavy traffic, use of studded tires, allowed velocity and information regarding the type of street. The input parameters are based on high quality yearly measurements. An example of how these measurements are carried out is given in an SLB report by Burman and Johansson (2010). There is also modelled data of emission, both from exhaust gases and in total, based on the modelled traffic data.

Modelled concentrations of PM_{10} are given from SIMAIR, and are divided into regional, urban, and local contributions. When summed up they can be compared to the measured concentration at Hornsgatan. The sum of the modelled regional and urban parts corresponds to the concentration in urban background, measured at Torkel Knutssonsgatan.

6.1.1 Characteristics of evaluation data

The time series is divided into an evaluation set and a validation set. The evaluation set is chosen as the data from 2007 and 2008, whilst the validation set consists of year 2009. Below are some characteristics of the evaluation set. No examination of the data from year 2009 is done, in order to not influence model choices.

Figure 5 shows a plot of the modelled and measured concentration of PM_{10} at Hornsgatan in 2007 and 2008.

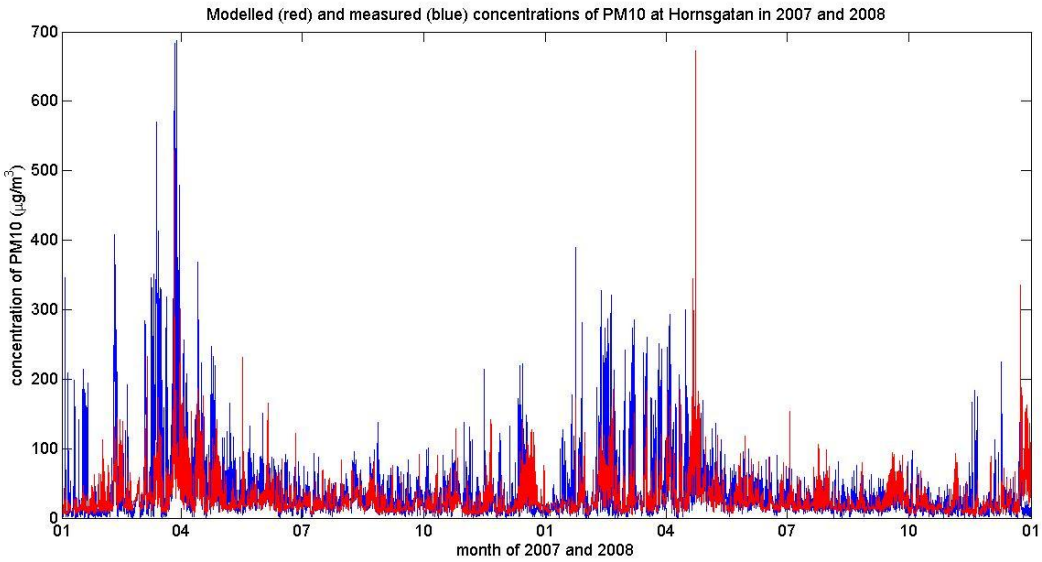

Figure 5. Concentration of PM_{10} at Hornsgatan in 2007 and 2008. Observations are seen in blue and the SIMAIR model is in red.

The concentration of PM_{10} is much higher during winter and early spring than during the rest of the year. One explanation is resuspended road dust, mainly caused by the use of studded tires during this period. The concentration peaks around March when the streets are drier and snow free, but studded tires are still in use. When these peeks occur, the model severely underestimates the concentration, while the model follows the observations well during the summer and early autumn.

The modelled and observed concentration of PM_{10} in urban background at the roof of Torkel Knutssonsgatan is illustrated in Figure 6.

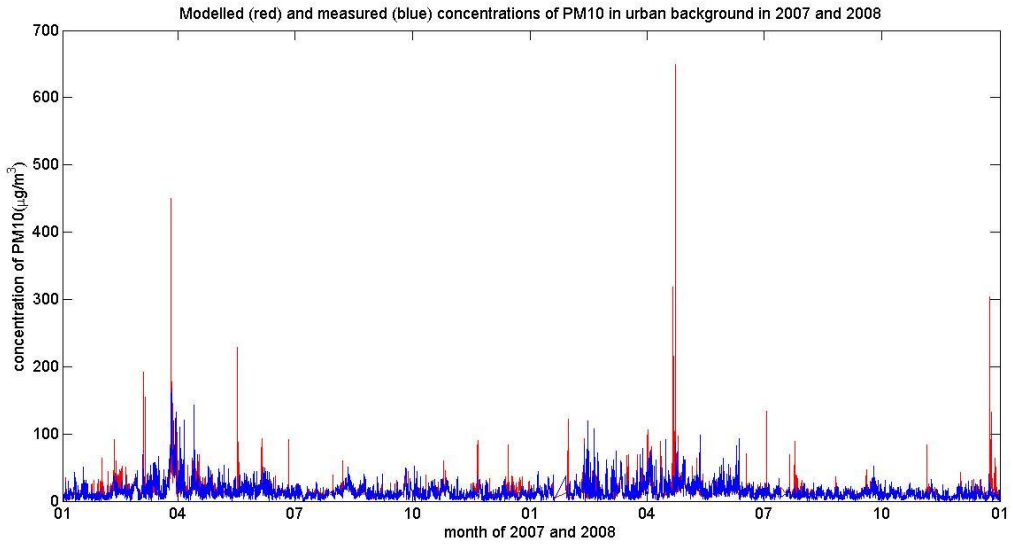

Figure 6. The concentration of PM_{10} in urban background at Torkel Knutssonsgatan. Observations are seen in blue and the SIMAIR model is in red.

There is a remarkable decrease in the concentration of PM_{10} 20 meters above ground compared to the streetscapes at Hornsgatan. Except for some false peaks the SIMAIR model performs very well for Torkel Knutssonsgatan.

Figure 7 shows the model error at Hornsgatan in 2007 and 2008, together with the model error mean value of 7.6 µg/m^3. The variance of the model error is higher during winter and early spring and lower during summer.

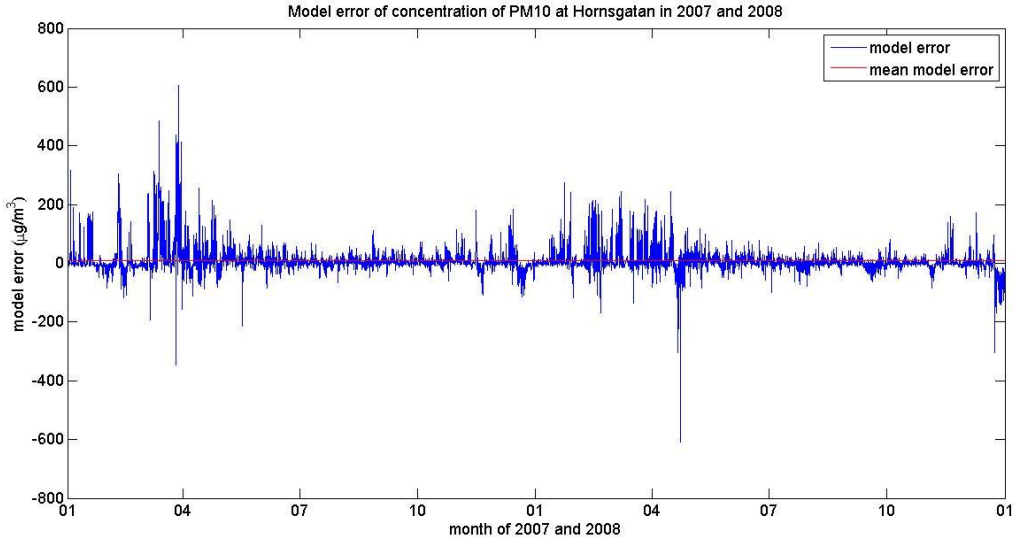

Figure 7. Model error (obs-mod) at Hornsgatan in 2007 and 2008, together with a red line indicating the model error mean value of 7.6 µg/m^3.

The normal probability plots for the model error and the error of the log-transformed model and observation data are provided in Figure 8. The model error does not belong to a normal

distribution. The transformed model error is closer to being normal distributed, but has a too heavy lower tail. This implies that a log-transform is not sufficient to obtain normal distributed errors, as required for e.g. linear regression analysis.

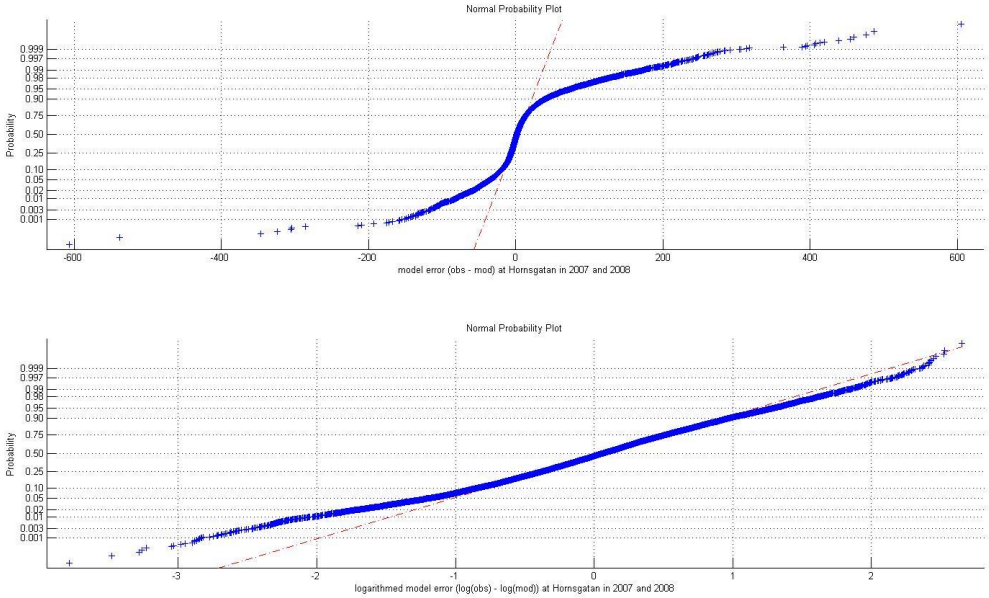

Figure 8. Normal probability plots for the model error on top and for the transformation $\log(obs) - \log(mod)$ below.

The SIMAIR model underestimates the concentration of PM_{10} for high measured values at Hornsgatan. The period of March 2007 and 2008 is examined more thoroughly, in order to see which properties that might be connected to high measurements. In Figure 9 the modelled and measured concentration of PM_{10} are seen for March 2007 and 2008. The model severely underestimates, but captures the daily variations. Around time point 600 very high concentrations are measured, and this time period will be examined more thoroughly later on.

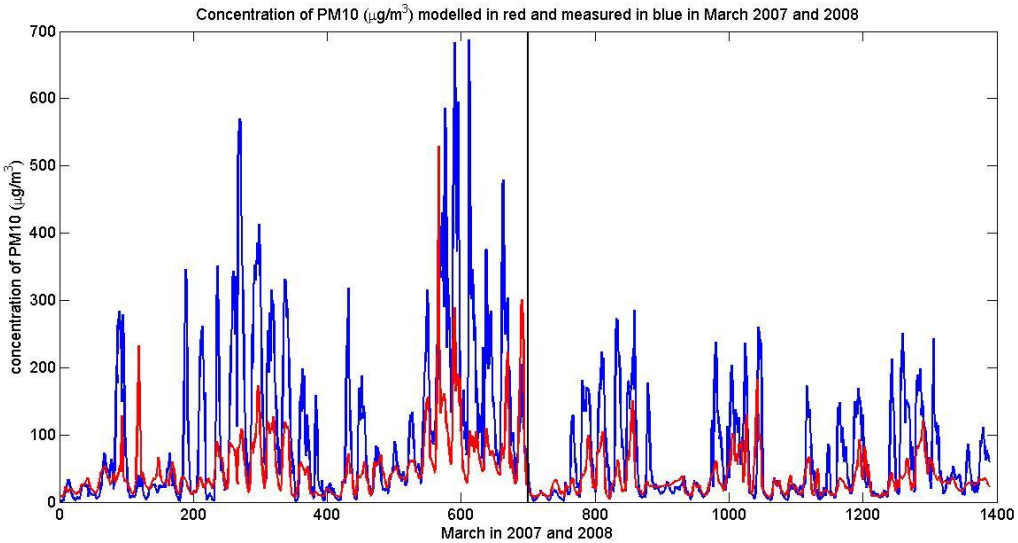

Figure 9. Hourly concentration of PM_{10} at Hornsgatan in March 2007 and 2008. Observations are seen in blue and the SIMAIR model is in red. The black line separates data from 2007 from 2008.

Some of the covariates are plotted for the same time period, in order to find potential relationships between covariates and increases in concentration. The subplots are shown in Figures 10 and 11.

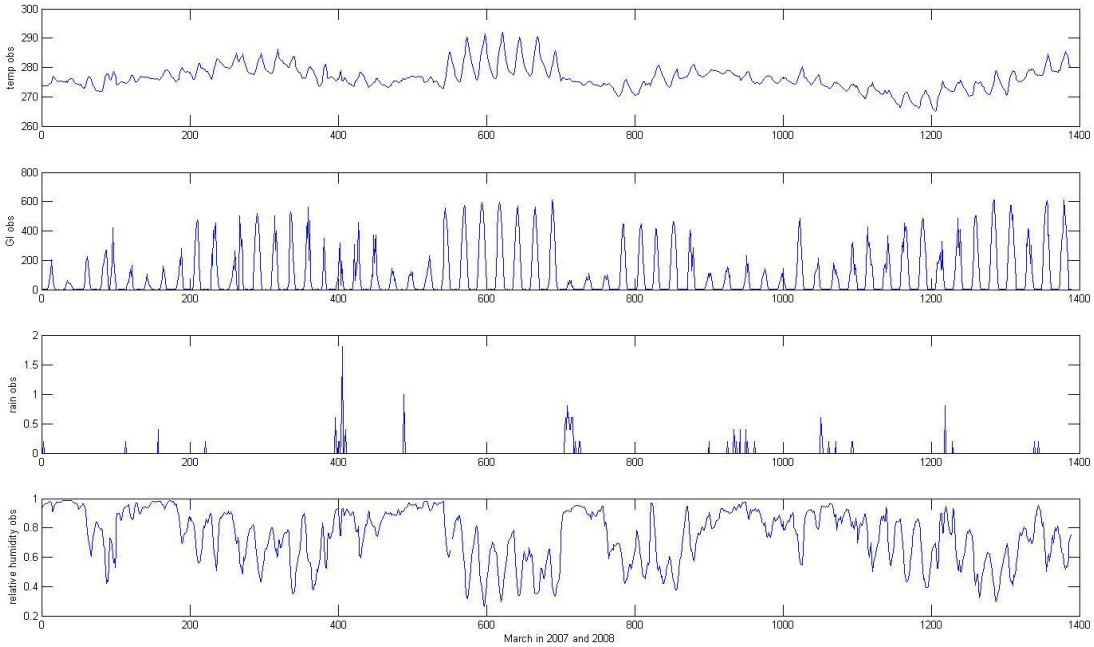

Figure 10. Hourly time series from top to bottom of measured (a) temperature in K, (b) global irradiance in W/m^2, (c) precipitation in mm and (d) relative humidity during March 2007 and 2008 at Hornsgatan.

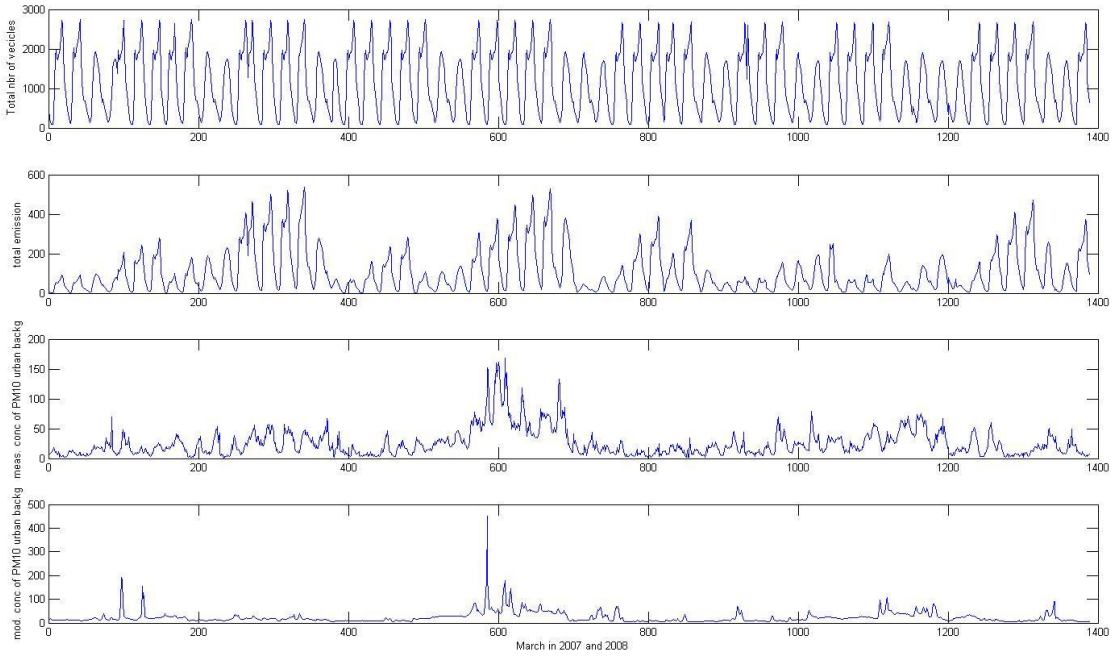

Figure 11. Hourly time series from top to bottom of (e) total number of vehicles, (f) total emission, (g) measured and (h) modelled concentration in µg/m^3 of PM_{10} in urban background during March 2007 and 2008 at Hornsgatan.

The first subplot (a) shows the measured temperature during March 2007 and 2008. Around time point 600 there are large variations in the daily minimum and maximum temperature. The second plot (b) shows the measured global irradiance, and the same pattern is seen here around time point 600. This implies a few sunny and warm days in a row. In the third plot (c) the measured precipitation is provided and the fourth plot (d) shows relative humidity. Around time point 600 there is no rain for several days. Plot number five (e) shows total number of vehicles; there is a clear weekly pattern with less traffic during weekends, and the days around time point 600 are week days. Plot six (f) shows modelled total emission. The factors outlined in plots (a) to (f) result in dry streets and many vehicles with studded tires around time point 600. This allows for high particle concentrations due to road wear dust, seen in Figure 9.

The last two plots are (g) measured and (h) modelled concentration of PM_{10} in urban background at Torkel Knutssonsgatan. Urban background concentration peaks occur around time point 600 as well.

One possible way of obtaining a better indication of when the streets are dry is to apply a weighted moving average filter for the time series of the precipitation. The exponentially weighted moving average used here has a time period of 18 hours. In Figure 12 the precipitation for March in 2007 and 2008 is shown without (a) and with (b) an exponentially weighted moving average.

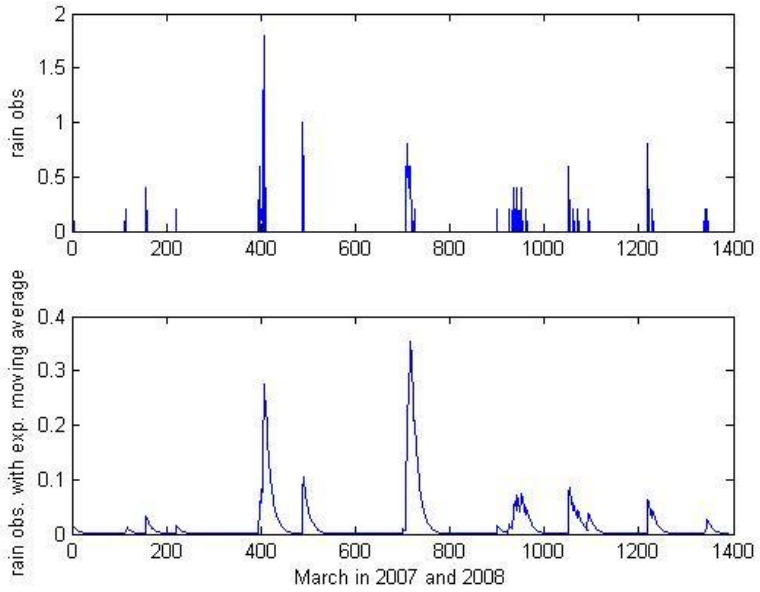

Figure 12. The hourly precipitation in March 2007 and 2008 at Hornsgatan, (a) without and (b) with an exponentially weighted moving average filter.

The smoothed rain time series can be used to construct an indicator for wet road surface. If the smoothed rain time series is lower than say $2.5 \cdot 10^{-3}$(threshold found by testing), the road surface is presumed to be dry, otherwise the street is considered wet. The result of this very simplified indicator is seen in Figure 13.

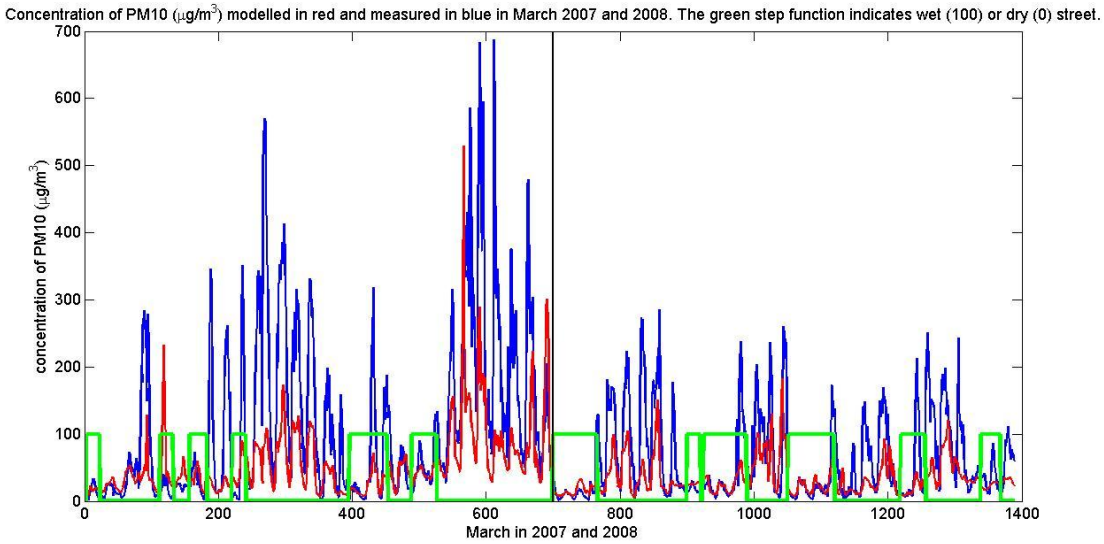

Figure 13. Concentration of PM_{10} at Hornsgatan in March 2007 and 2008. Observations are seen in blue and the SIMAIR model in red. A step function in green illustrates conditions for wet (100) and dry (0) streets based on the exponentially smoothed precipitation time series. The black line separates data from 2007 from 2008.

The step function in green gives an indication of conditions that are suitable for very high concentrations. It can be noted that the longer a dry period lasts, the higher the concentration of PM_{10} becomes. However, it is important to remember that there are many factors that affect how fast the street dries after a rainfall.

An approach to investigating patterns related to high peaks is to choose a threshold of say 100 µg/m^3. All observations above this threshold are collected in a new dataset and evaluated, together with model values for the same indices. With the threshold set to 100 µg/m^3 the following results were obtained. In Table 3 the mean absolute error (MAE) and root mean square error (RMSE) are calculated for all data, for peaks and for non-peaks.

Table 3. Mean Absolute Error and Root Mean Square Error for the model error at Hornsgatan in 2007 and 2008, for all data, for peaks and for non-peaks.

MAE of all data	MAE of obs. points > 100	MAE of obs. points < 100	RMSE of all data	RMSE of obs. points > 100	RMSE of obs. points < 100
21.21	107.23	14.70	41.27	127.76	24.42

There are 1192 points above the threshold and 15748 points below in the evaluation set. It is clear that the model needs to be improved when it comes to dealing with high concentrations. In Figure 14 histograms of the measured concentration above the observation threshold and the modelled values for the same indices of PM_{10} are shown. There is a clear bias between the model and the observations.

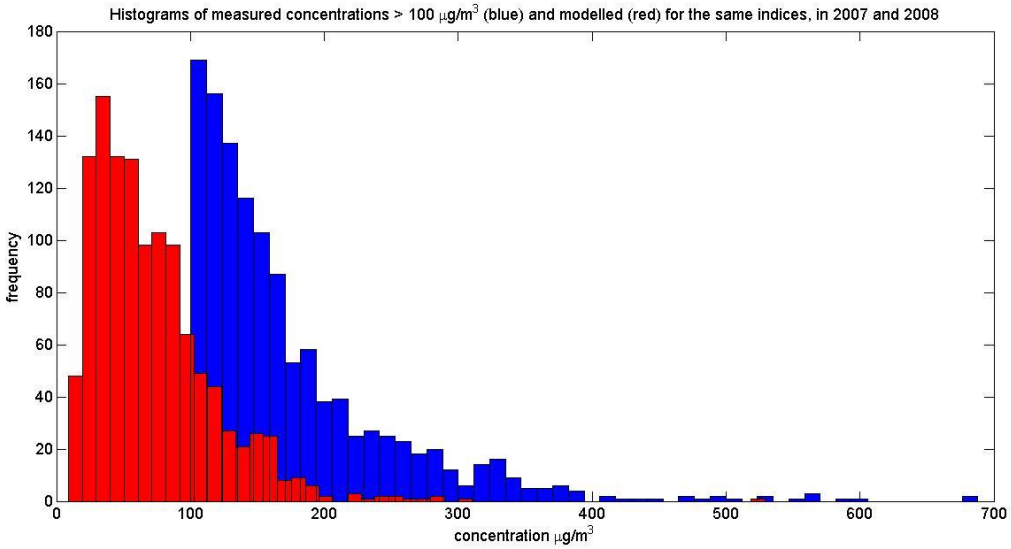

Figure 14. Histograms of modelled (red) and measured (blue) concentration of PM_{10} for time points in 2007 and 2008 where measurements exceed 100 µg/m^3.

In order to improve the model performance on concentration peaks it is important to know where and why the peaks occur. The indices of the top 1192 points for the observations are compared to the top indices for the model. Only 41 % of the model indices agree with the observation indices. This implies that more than half of the model maxima do not occur at the same time points as measurement maxima.

In Figure 15 the number of peaks that occur during each hour of the day is plotted. The majority of the peaks occur during day time, probably primarily due to traffic intensity.

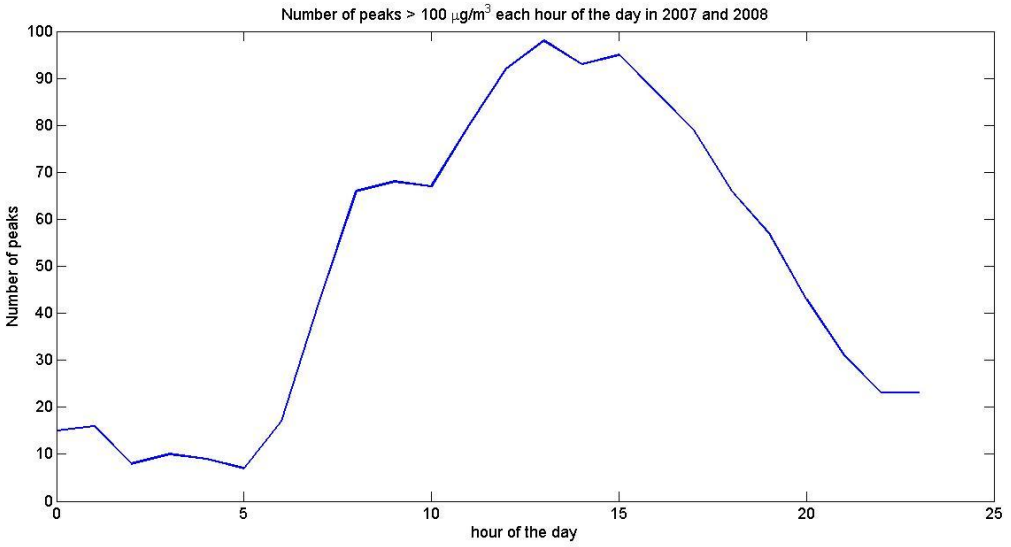

Figure 15. The number of times per hour of the day that measurements of PM_{10} exceed 100 µg/m^3 at Hornsgatan in 2007 and 2008.

Due to this tendency each day is divided into day (07.00 through 20.00 UTC) and night (21.00 through 06.00 UTC), where UTC stands for Coordinated Universal Time. The day/ night indicators are used as input to the SVR, since indicators for each hour of the day would increase the risk of overfitting.

In Table 4 the number of observed peaks above 100 µg/m^3 during each month of the year in 2007 and 2008 are listed.

Table 4. Number of measurements above 100 µg/m^3 in each month of the year of 2007 and 2008 at Hornsgatan.

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Nbr of peaks	117	206	424	299	18	1	0	7	0	7	44	69

Based on the pattern in this table each year is divided into three seasons – winter (November, December and January), spring (February, March, April) and summer (May through October).

The correlation between model error and different covariates provides a way of testing if a covariate contains useful information that can be used to correct the model. It is interesting to look at both linear correlation (Pearson's r) and if the relation can be described as a monotone function (Spearman's rank). The linear correlation between model error and each covariate in the test data from 2007 and 2008 are seen in Tables 5, 6 and 7. The data is divided into the seasons mentioned above. Where no value is given there is no significant correlation according to a significance test using the Student's t-distribution in Matlab.

Table 5. Significant linear correlation (Pearson's r) between the model error and humidity, precipitation and GI, in 2007 and 2008 at Hornsgatan.

	Humidity obs.	Humidity MESAN	Precipitation obs. (exp. filtered)	Precipitation MESAN (exp. filtered)	GI obs.	GI STRÅNG
2007 - 2008 winter period:	-0.39	-0.41	-	-	0.10	0.12
2007 - 2008 spring period:	-0.18	-0.20	-0.06	-0.09	0.14	0.13
2007 - 2008 summer period:	0.13	0.13	-	0.04	-0.07	-

Table 6. Significant linear correlation (Pearson's r) between the model error and concentration in urban background, number of vehicles (in total, heavy and light) and total emission, in 2007 and 2008 at Hornsgatan.

	PM10 urban background obs.	PM10 urban background mod.	Nbr of vehicles	Nbr of heavy vehicles	Nbr of light vehicles	Total emission
2007 - 2008 winter period:	0.33	-0.34	0.08	0.11	0.07	-0.22
2007 - 2008 spring period:	0.54	-0.16	0.21	0.23	0.20	0.19
2007 - 2008 summer period:	0.31	-0.24	-	0.14	-	-0.20

Table 7. Significant linear correlation (Pearson's r) between the model error and temperature, u- and v-component of wind and wind direction, in 2007 and 2008 at Hornsgatan.

	Temp. obs.	Temp. MESAN	u-wind obs.	u-wind MESAN	v-wind obs.	v-wind MESAN	Wind dir. Obs.	Wind dir. MESAN
2007 - 2008 winter period:	0.10	0.10	-	-	0.09	0.09	-	-
2007 - 2008 spring period:	0.07	0.09	-	-	0.16	0.16	0.15	0.13
2007 - 2008 summer period:	-0.14	-0.14	-0.06	-0.11	0.09	0.08	0.08	0.05

Several covariates display large differences in their correlation with model errors for the different seasons, some even change signs. It can be noted that the correlation between model error and meteorological parameters does not differ a lot between measurements and MESAN data.

The corresponding table for correlation using Spearman's rank can be seen in Appendix B. For example the exponentially smoothed precipitation shows a higher correlation with the model error using Spearman's rank instead of Pearson's r.

These tables of correlation function as guidance when deciding which parameters are to be used in the Support Vector Regression.

6.2 Validation data

Data from Umeå and Gothenburg during 2007 to 2009 are used for validation, in order to test the generality of the statistical model. The characteristics of the data are therefore not examined, and the available data are only briefly discussed.

6.2.1 Umeå

In Umeå there is an observation site at the east side of Västra Esplanaden, a part of the E4 highway that goes through central parts of Umeå. Approximately 24 000 vehicles pass here every day, and 8 % are heavy vehicles. This streetscape site, under care of the Transport Administration, measures hourly concentrations of PM_{10} and NO_2. The street is 28 meters wide and there are 15 meters high buildings on both sides. The speed limit is 50 km/h. Sand is used when lanes are slippery. During winter 88 % of the vehicles used studded tires in 2007, 83 % in 2008 and 94 % in 2009.

There is an observation site on the roof of the library in central Umeå that measures hourly concentrations of pollutants in urban background. This station is about 400 meters away from the station at Västra Esplanaden and measurements are carried out by Umeå municipality (Andersson S. and Omstedt G. 2012, p. 5-7).

In Figure 16 the observation site at Västra Esplanaden is marked with a blue dot, and the station at the library is marked with a red dot.

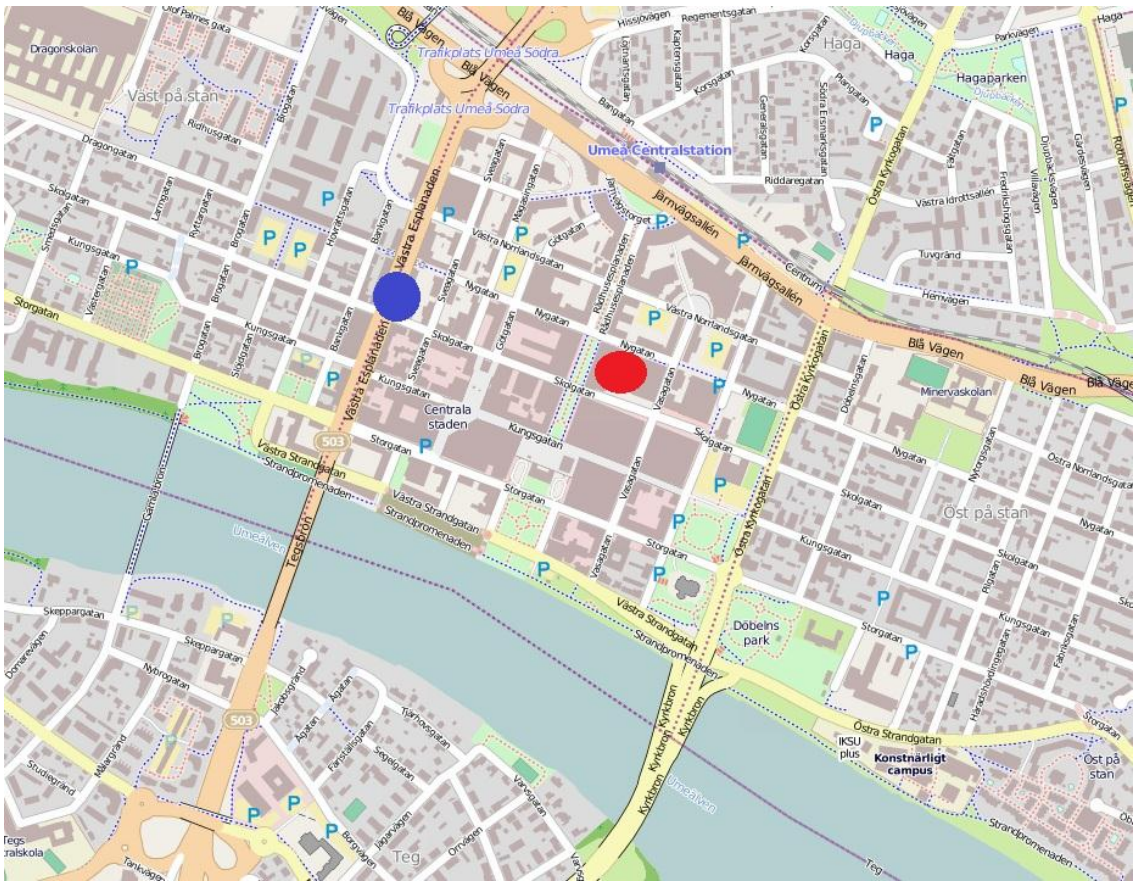

Figure 16. The observation site at Västra Esplanaden (streetscape) is marked with a blue dot and the station at the library (urban background) is marked in red. Source: OpenStreetMap.

6.2.2 Gothenburg

There is an observation site at the E6 highway at Gårda in Gothenburg. There are three lanes in each direction and the yearly daily number of vehicles is around 90 000, of which about 7 % is heavy traffic. On the west side of the highway there is a rock wall and some houses which are about 7 meters high. On the other side there are 10 meters high buildings, and the street is 64 meters wide in total. The observation site is situated on the west side and measures hourly pollutant concentrations. The speed limit is 70 km/h and salt is used in winter conditions. During winter 74 % of the vehicles used studded tires in 2007, 71 % in 2008 and 68 % in 2009.

There is another observation site at the roof of the Femman building in central Gothenburg. This station measures concentrations of PM_{10} and NO_2 in urban background on an hourly basis. Both observation sites are handled by the Environmental Department at Gothenburg municipality (Andersson S. and Omstedt G. 2012, p. 2-3).

In Figure 17 the observation site at Gårda is marked with a blue dot, and the Femman building is marked with a red dot.

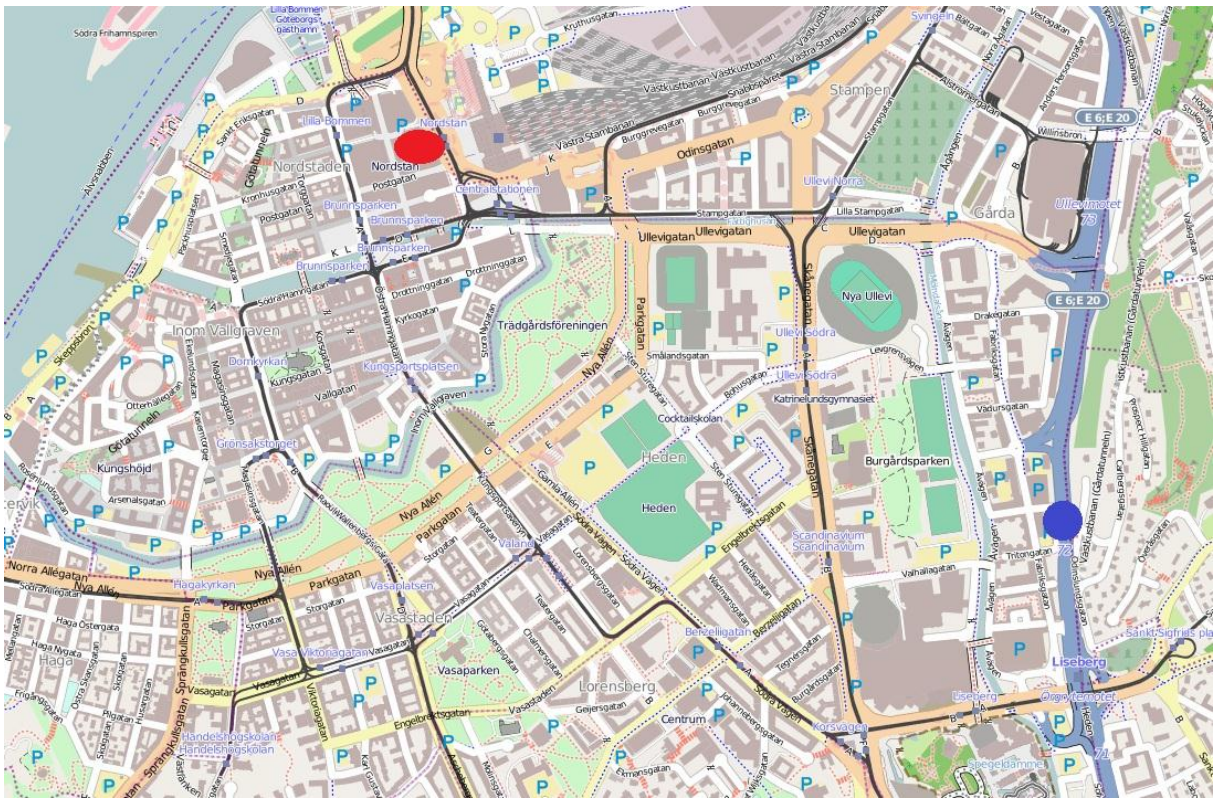

Figure 17. The observation site at Gårda (streetscape) is marked in blue and the station at the Femman building (urban background) is marked with a red dot. Source: OpenStreetMap.

7 Results

As described in Section 6 the data used to develop a statistical post-processing model are from Hornsgatan in Stockholm. To test generality, the resulting model is validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg.

The results from the Support Vector Regression for the three places are presented, and for Hornsgatan the same method is also tested with meteorological observations as input. A comparison of relevant meteorological parameters measured in Stockholm, with data from MESAN and STRÅNG is also done.

The resulting C and γ found by validation for each of the three regressions described in Section 4.5, for all sites, are given in Appendix C.

7.1 Stockholm

7.1.1 Support Vector Regression using MESAN and STRÅNG parameters

The explanatory variables used in the Support Vector Regression for Hornsgatan data are given in Table 8. All meteorological parameters are from MESAN, except for global irradiance which has been taken from STRÅNG. In addition to these parameters there are also indicators for day or night and for winter, spring or summer.

Table 8. The explanatory variables used in the Support Vector Regression of data from Hornsgatan.

SIMAIR model	SIMAIR urban background	Nbr of heavy vehicles	Precipitation (exp. filtered)	Relative humidity	Maximum temp. difference each day	Acc. GI each day	Wind direction

These variables are chosen based on discussions with the SIMAIR researchers, by studying correlation tables and by testing different combinations of covariates.

Table 9 gives an overview of the results of the Support Vector Regression compared to the SIMAIR model for Hornsgatan in 2009.

Table 9. A summary of the results of the Support Vector Regression compared to the SIMAIR model at Hornsgatan in 2009.

	Observations	SIMAIR model	SVR
Yearly mean (µg/m^3)	37.2	28.2	34.1
90-percentile daily mean (µg/m^3)	81.9	59.3	67.0
Days > 50 µg/m^3	67	46	54
RPE %		24	8.3
RDE %		22	7.7
r daily mean		0.73	0.80
r hourly		0.63	0.68
RMSE daily mean (µg/m^3)		14.2	12.4
RSME hourly (µg/m^3)		36.0	33.1

There are improvements in all comparisons.

In Figures 18 and 19 the observations at Hornsgatan are plotted together with the SIMAIR values and the Support Vector Regression, on hourly and daily basis.

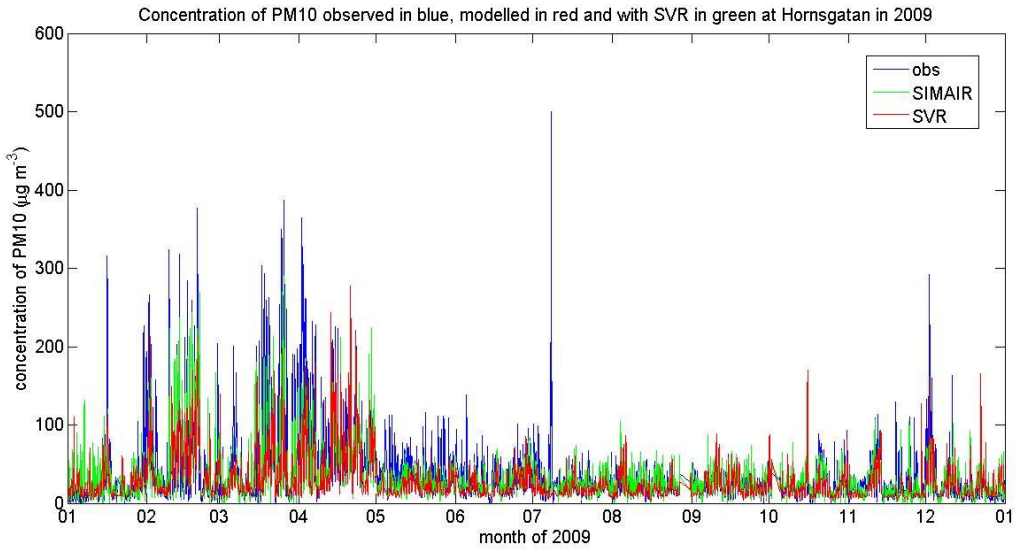

Figure 18. Measurements of PM_{10} at Hornsgatan in 2009 are in blue, SIMAIR model values in red and SVR values in green.

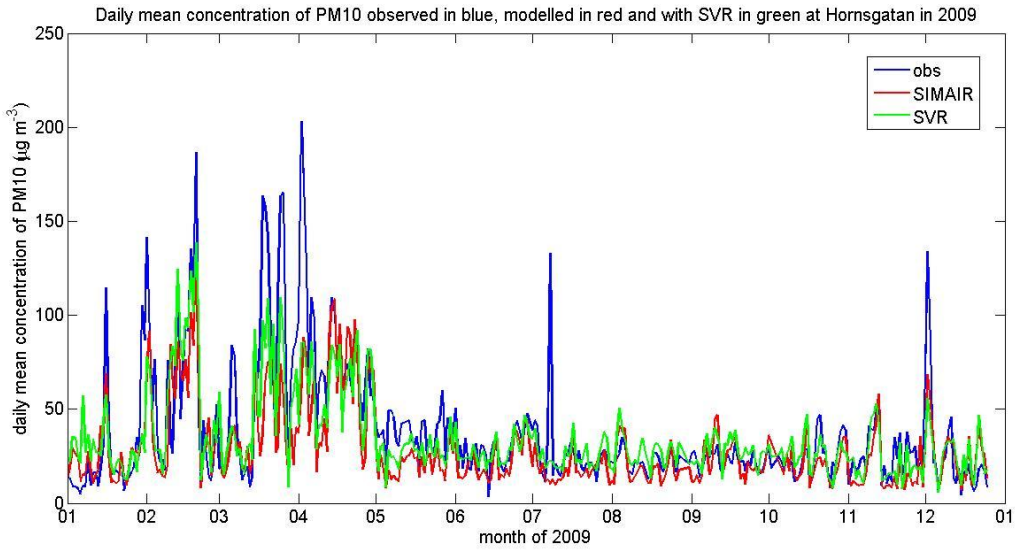

Figure 19. Daily average of measurements of PM_{10} at Hornsgatan in 2009 in blue, of SIMAIR model values in red and of SVR values in green.

Some improvements are seen, however the highest peaks are still difficult to correct.

Figure 20 shows a log-log plot of the percentiles for the models and the observations. The red line compares the percentiles of SIMAIR and measurements, whilst the green line compares the percentiles of the SVR and measurements. The black line shows a perfect match.

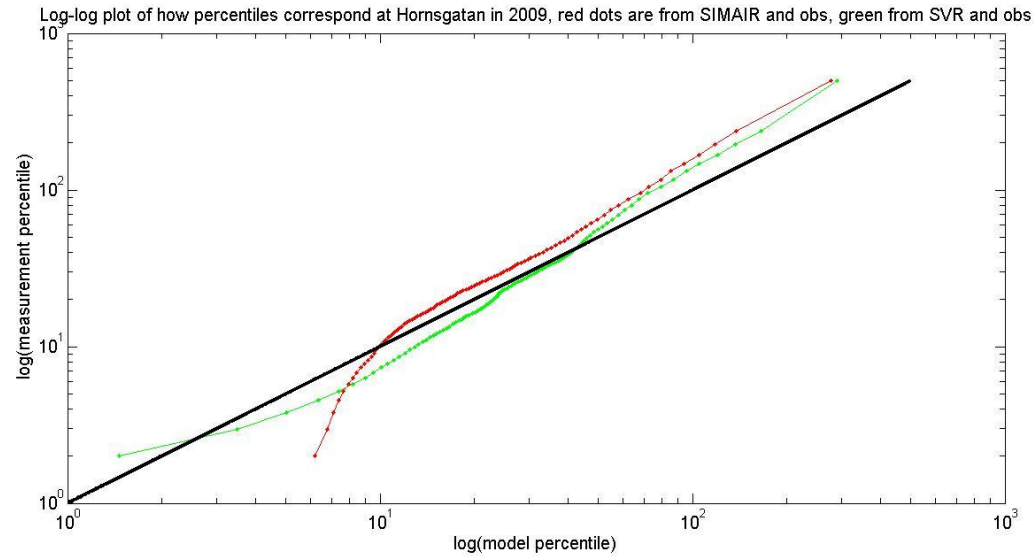

Figure 20. A log-log plot of the percentiles for the models and the observations at Hornsgatan in 2009. The red line compares the percentiles of SIMAIR and measurements and the green line compares the percentiles of the SVR and measurements. The black line shows a perfect match between model and measurement percentiles.

For almost all percentiles the SVR-corrected SIMAIR outperforms the SIMAIR model.

7.1.2 Support Vector Regression using meteorological observations

To see if any of the meteorological observations at Torkel Knutssonsgatan can better capture the changes in concentration of PM_{10} than the meteorological parameters from MESAN, the MESAN data is replaced by observations in the Support Vector Regression.

After a lot of testing the only parameter that somewhat improves the SVR is the observed wind direction. The reason why the observed wind direction gives slightly more information is probably that the terrain affects wind a lot. The wind effect on PM_{10} concentration is captured better using measurements nearby than a more smoothed gridded data from MESAN. The fact that no larger deviations are found suggests that the method is insensitive to the use of MESAN or observational data.

The table of results using the same method as described in Section 7.1.1 with the difference that MESAN wind direction is replaced by observed wind direction is seen in Appendix D. The results are very close to the results presented in Section 7.1.1, but with a minor decrease in both RMSE and (unfortunately) yearly mean value.

7.1.2.1 Comparison between meteorological observations and MESAN/ STRÅNG

As there is a slight improvement in the Support Vector Regression when using wind direction observations from Torkel Knutssonsgatan instead of data from MESAN, it is interesting to examine the wind direction residual (where the residual is the difference between measured wind direction and MESAN data). The histogram of the wind direction residual is plotted in Figure 21, and in Figure 22 a normal probability plot of the residual is given.

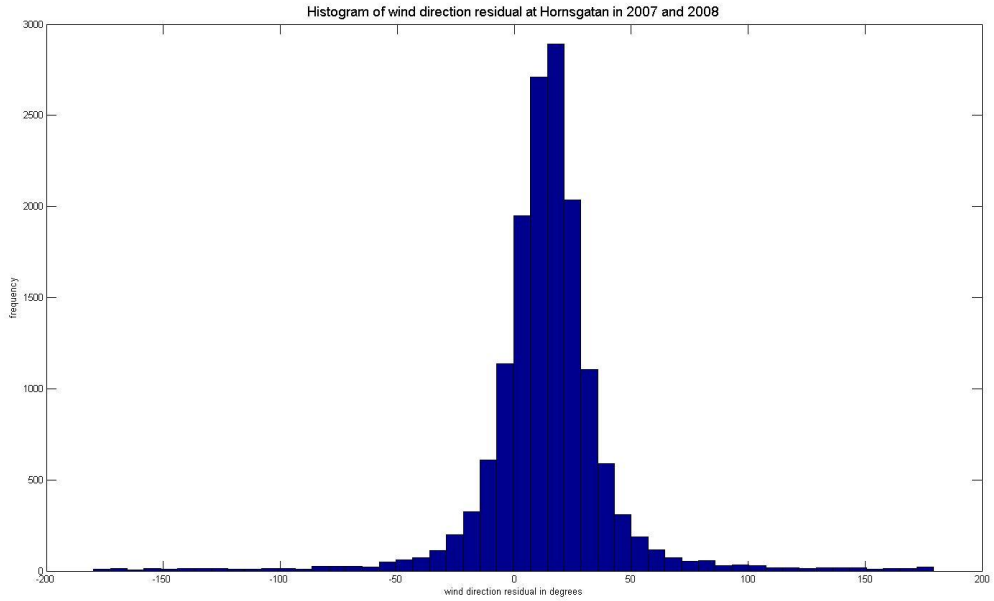

Figure 21. A histogram of the wind direction residual at Hornsgatan in 2007 and 2008.

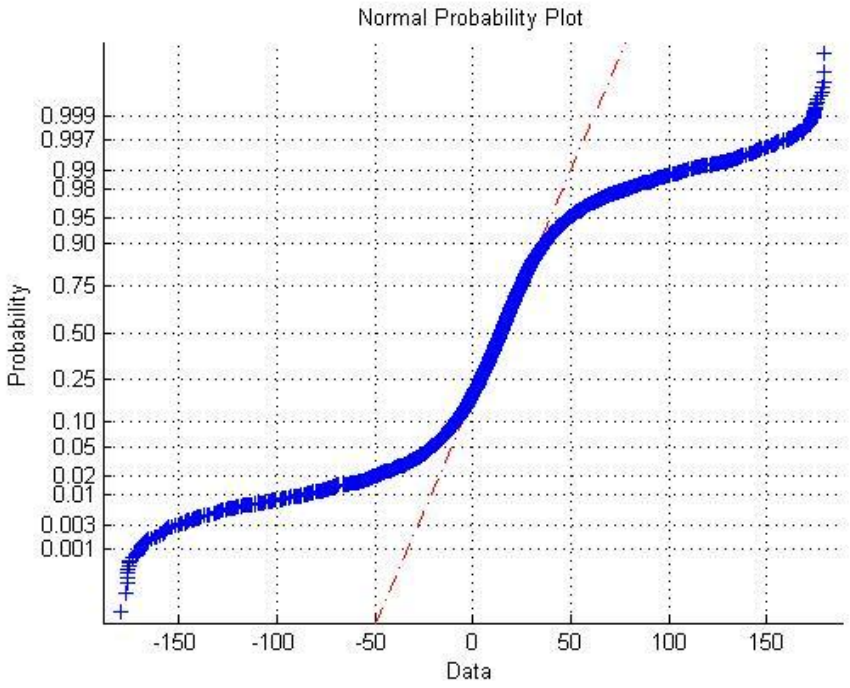

Figure 22. A normal probability plot of the wind direction residual at Hornsgatan in 2007 and 2008. A good fit to a normal distribution would lie close to the red line.

The residual has a mean of 13.9 degrees. The residual is not normal distributed, which is seen in the normal probability plot. The residual has heavier tails and it is limited to the interval [-180, 180].

7.2 Validation results

The statistical post-processing model is validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg. The SVR model uses the same explanatory variables and methodology as for Hornsgatan, with the addition of wind speed at Gårda. Training and setting of parameters C and γ is done using data from the specific location. The results are promising for both sites, especially in correcting yearly mean value. This implies that the SVR model generalizes well.

7.2.1 Umeå

7.2.1.1 Support Vector Regression using MESAN and STRÅNG parameters

Table 10 shows a summary of the results of the Support Vector Regression compared to the SIMAIR model for Västra Esplanaden in 2009.

Table 10. A summary of the results of the Support Vector Regression compared to the SIMAIR model for Västra Esplanaden in 2009.

	Observations	SIMAIR model	SVR
Yearly mean (µg/m^3)	22.2	28.7	20.9
90-percentile daily mean (µg/m^3)	45.8	55.4	36.4
Days > 50 µg/m^3	33	45	21
RPE %		29	6.0
RDE %		16	3.4
r daily mean		0.45	0.54
r hourly		0.36	0.40
RMSE daily mean (µg/m^3)		15.8	11.2
RSME hourly (µg/m^3)		35.5	32.4

In Figures 23 and 24 the measurements of PM_{10}, on hourly basis and as daily mean, are plotted together with SIMAIR and SVR values.

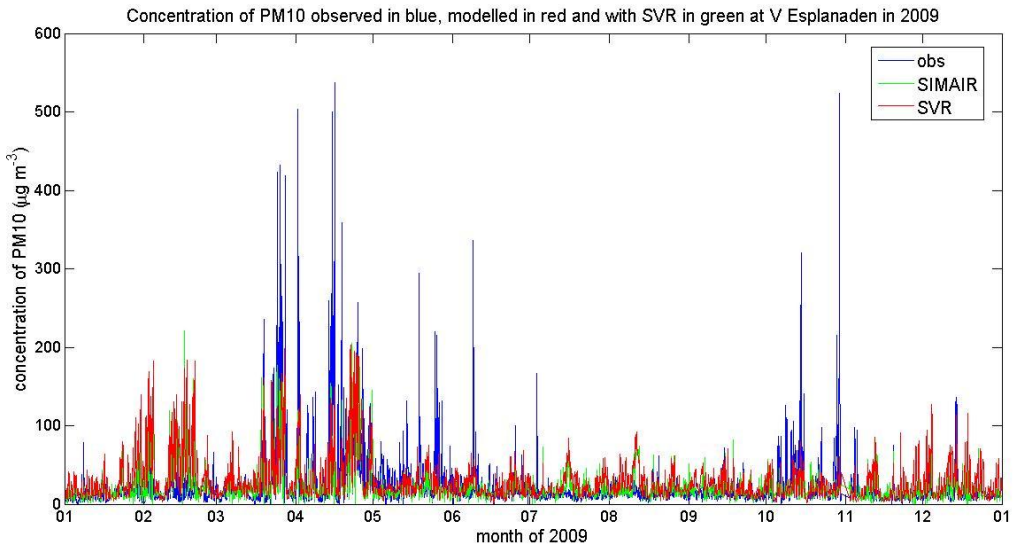

Figure 23. Measurements of PM_{10} at Västra Esplanaden in 2009 in blue, SIMAIR model values in red and SVR values are plotted in green.

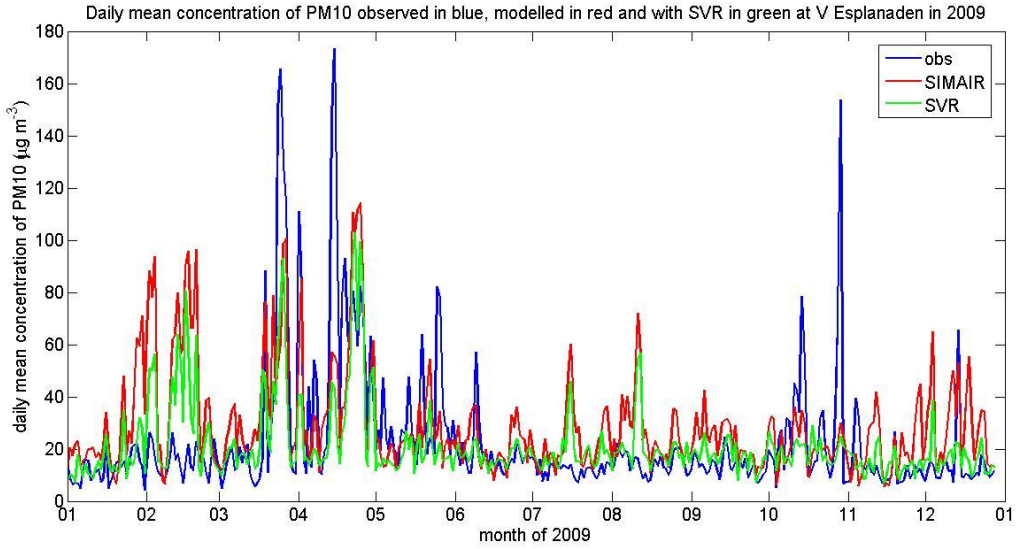

Figure 24. Daily average of measurements of PM_{10} at Västra Esplanaden in 2009 in blue, SIMAIR model values in red and SVR values are in green.

Figure 25 shows a log-log plot of the percentiles for the models and the observations.

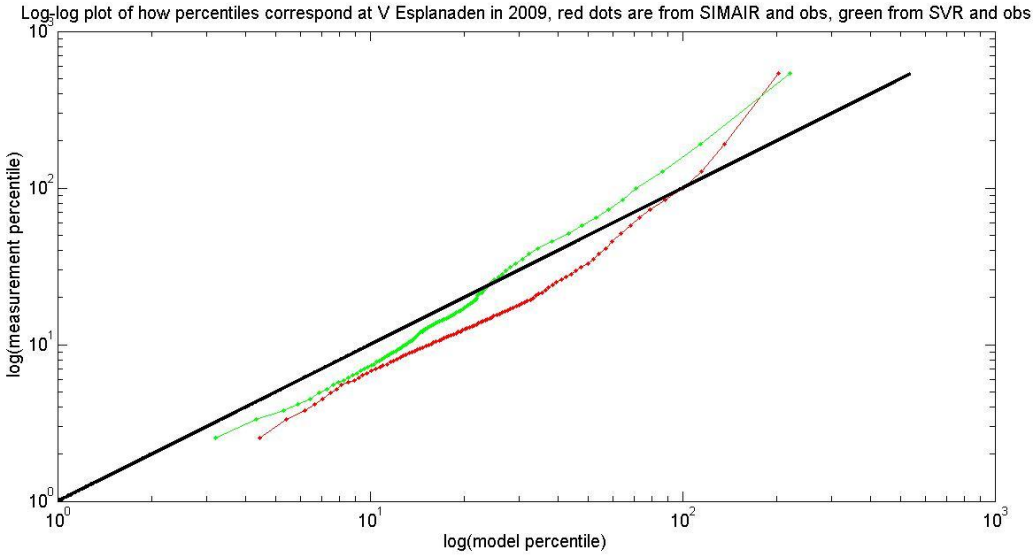

Figure 25. A log-log plot of the percentiles for the models and the observations at Västra Esplanaden in 2009. The red line compares the percentiles of SIMAIR and measurements and the green line compares the percentiles of the SVR and measurements. The black line shows a perfect match between model and measurement percentiles.

It can be noted that for almost all percentiles the regression performs better than the SIMAIR model. For some of the higher percentiles the SVR model underestimates while the SIMAIR model overestimates. The highest peaks are underestimated both by the SVR modified model and the raw SIMAIR model.

7.2.2 Gothenburg

7.2.2.1 Support Vector Regression using MESAN and STRÅNG parameters

The same explanatory variables and methodology as for Hornsgatan are used for Gårda, with the addition of wind speed. The reason for including wind speed is that the result improves a bit. Probably the wind speed contributes here but not at Hornsgatan or at Västra Esplanaden because the streetscape is much more open at Gårda and because wind speeds are generally higher in Gothenburg than in Stockholm and Umeå.

In Table 11 the results for Gårda in 2009 are summarized.

Table 11. A summary of the results of the Support Vector Regression compared to the SIMAIR model at Gårda in 2009.

	Observations	SIMAIR model	SVR
Yearly mean (µg/m^3)	23.7	37.0	24.8
90-percentile daily mean (µg/m^3)	39.3	70.3	39.3
Days > 50 µg/m^3	15	68	15
RPE %		56	5.6
RDE %		33	3.0
r daily mean		0.60	0.76
r hourly		0.43	0.57
RMSE daily mean (µg/m^3)		15.0	6.49
RSME hourly (µg/m^3)		29.5	18.3

In Figures 26 and 27 the measurements of PM_{10} are plotted, together with SIMAIR and SVR values, on hourly basis and as daily mean. There are missing data in October 2009.

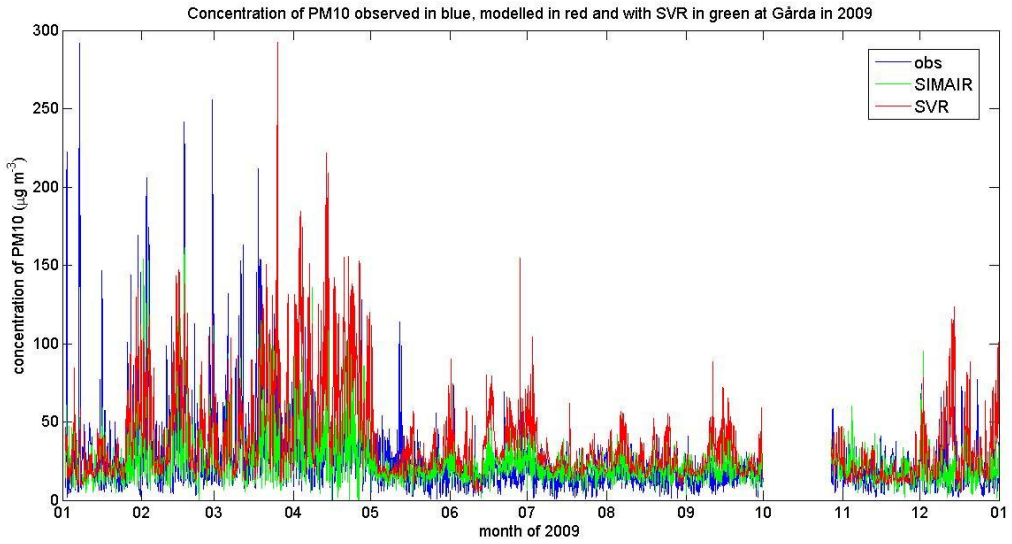

Figure 26. Measurements of PM_{10} at Gårda in 2009 in blue, SIMAIR model values in red and SVR values are in green.

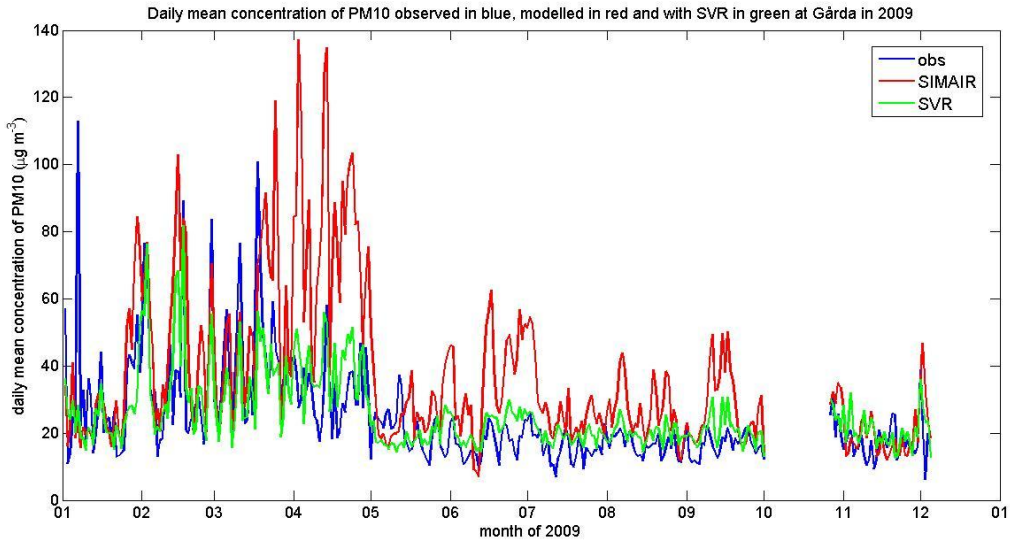

Figure 27. Daily average of measurements of PM_{10} at Gårda in 2009 in blue, SIMAIR model values in red and SVR values are in green.

Especially the daily mean concentration improve significantly, as the overestimation in the raw SIMAIR model is corrected by the SVR-modified model.

A log-log plot of the percentiles for the models and the observations is provided in Figure 28.

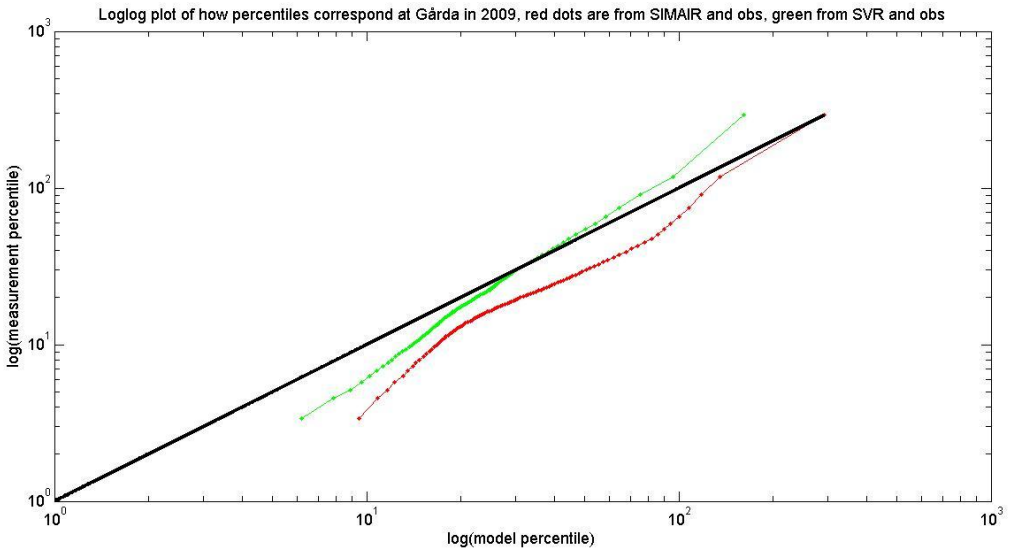

Figure 28. A log-log plot of the percentiles for the models and the observations at Gårda in 2009. The red line compares the percentiles of SIMAIR and measurements and the green line compares the percentiles of the SVR and measurements. The black line shows a perfect match between model and measurement percentiles.

The SVR percentiles correspond better to measurements for all but the highest few percentiles. The 90-percentile of the SVR-modified model is a perfect fit to measurements, which can also be seen in Table 11.

7.3 Target diagram

In Figure 29 the Target diagram, produced by Delta Tool, is given for the three sites in 2009 with and without statistical post-processing. As the points are on the left hand side the errors are dominated by low correlation as opposed to large standard deviation errors. For all sites both bias (the deviation between modelled and measured yearly mean values) and centred RMSE decrease with SVR. For Gårda and Hornsgatan the Target decrease from >1 to <1, which indicates that the statistically post-processed model values might describe the concentration of PM_{10} better than the measurements.

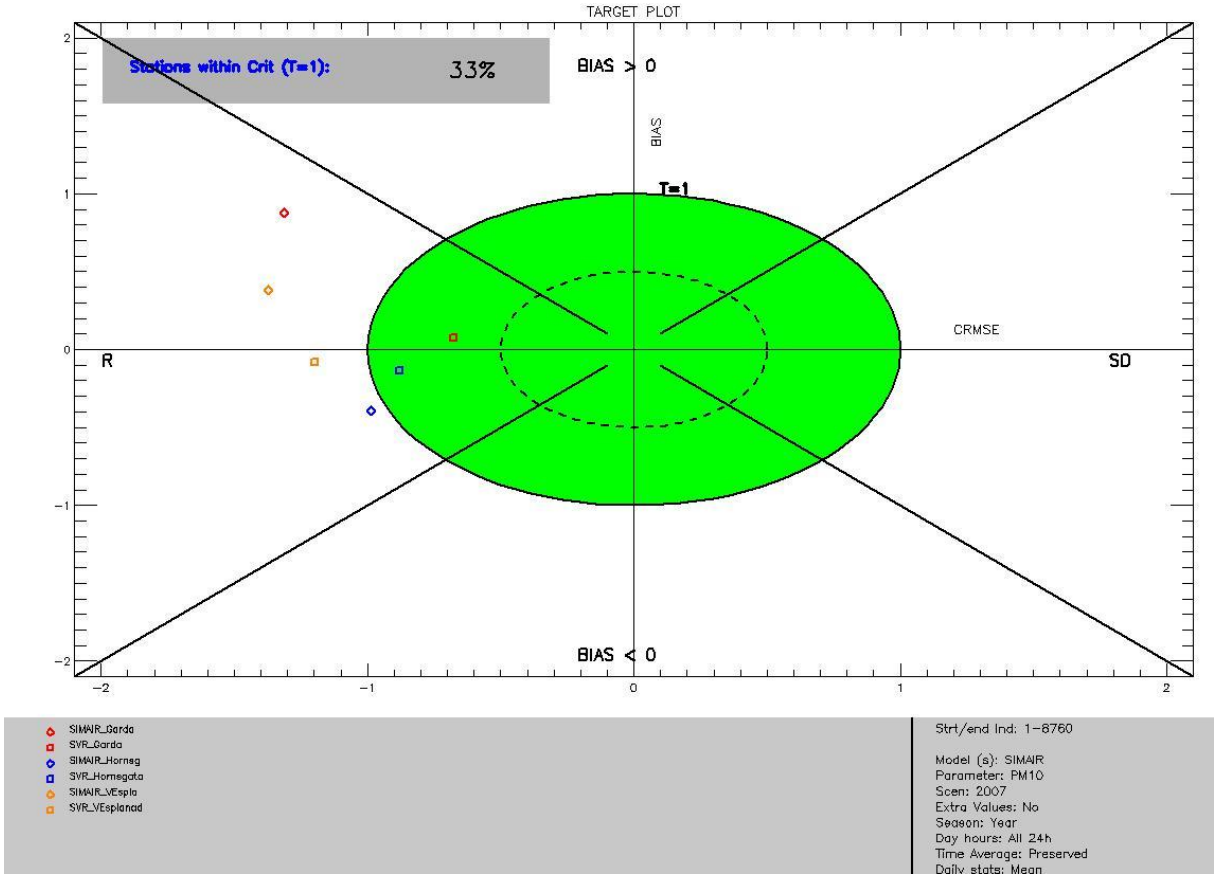

Figure 29. A Target plot showing the performance of the SIMAIR model for Hornsgatan with blue diamonds, Gårda with red diamonds and Västra Esplanaden with orange diamonds. The squares visualize how the SVR-modified model influences the performance for the three sites. For Hornsgatan and Gårda the Target is decreased to be <1 with SVR. Target is also decreased for Västra Esplanaden but is still >1.

8 Discussion and conclusions

Thanks to research and international cooperation, the area of combining monitoring data and air quality models is developing fast. Statistical methods can contribute substantially to air quality modelling by integrating information from measurements.

At SMHI Support Vector Regression has been successfully implemented in some products and this thesis shows promising results within air quality applications. There are several advantages with this method. It is simple to implement, with only a few crucial parameters to tune. SVR can handle non-linarites well and does not suffer very much from overfitting. The method is computationally efficient since only part of the input data (the support vectors) are used to find the optimal hyperplane. However, it is important to remember that similar behaviour of the training data and test data is needed in for the SVR to perform well. For example the ban on studded tires at Hornsgatan in 2010 makes it unreasonable to use the SVR on data from 2010 based on previous years as training data. Another disadvantage of the SVR is the difficulty in interpreting how the input parameters affect the output, as the output (in this case) consists of a linear combination of Radial Basis Functions.

Other statistical methods, such as Kalman based methods, may perform as well and provide an easier physical interpretation, but a lot more effort is needed to choose appropriate parameters.

The explanatory parameters for the SVR need to be selected carefully to achieve a good performance. As in the case with other regressions the covariates should not be (too) correlated with each other. The covariates in this thesis were chosen by studying correlation tables, talking to the SIMAIR researchers Stefan Andersson and Gunnar Omstedt and by testing different combinations of inputs.

The hold out-validation was implemented since it was computationally less expensive than a k-fold cross validation. The data set is quite large so by randomly picking 25% of the evaluation data as validation, hopefully not too much information is lost. However, a k-fold cross validation may be more appropriate for smaller data sets.

The best parameters C and γ in the grid search are found based on smallest MAE of the validation set. Using the RMSE to decide the parameters instead, led to basically the same results for Stockholm and Gothenburg, but for Umeå C increased a lot and the final result was almost as bad as the original model. Probably this was due to the fact that RMSE penalizes large deviations more than MAE does, leading to overfitting. A higher C aims at finding a hyperplane that fits all training samples well.

The choice of implementing three regressions, depending on the value of the SIMAIR model, was a way of easing the fitting of hyperplanes. At Hornsgatan the SIMAIR model severely underestimated the peaks, so by separating peak values, medium and low values better regression performance was hoped for. In a comparison with just one regression three regressions gave a slightly lower RMSE for Hornsgatan data. Future studies could examine if there are better ways to divide the data and what number of regressions would give superior results.

The reason the test data were set to be a whole year, and not for example a randomly picked part of the three years of the available data, was that the SIMAIR model is run one year at a time in production mode. This structure will facilitate the integration of the program with the regular simulation. An interesting question is if a longer training period will improve results.

The SVR performed well on all three sites, with improvements in basically all statistical indicators. The yearly mean value is an indicator that improved greatly, which resulted in large decreases in RPE and RDE. It was a bit unexpected that out of the three sites there was

least improvement (for certain indicators) on the Hornsgatan data, even though that data set was used to construct the statistical model. One possible explanation is that the emission model for road dust used in SIMAIR was adjusted to fit Hornsgatan data, and was then generalized to other places. Another reason could be that the concentration peaks at Hornsgatan were severely underestimated, but otherwise the SIMAIR model fitted the measurements well. The peaks were problematic to correct, maybe because they were difficult to predict due to lack of important data. Activities such as salting and gritting in winter conditions or cleaning of the streets have great impact on particle concentrations but there are no available data regarding these activities.

The greatest improvements were seen in the SVR for Gårda data, where the SIMAIR model overestimated all percentiles. The structure of the streetscapes at Gårda is complicated with open surroundings on one side of the street. The air quality modelling researchers tried to use both the OSPM and OpenRoad dispersion models and decided on OSPM, though neither gave satisfying results. Gårda is heavily trafficked; more than three times more vehicles pass there compared to Hornsgatan and Västra Esplanaden on an average day. The percentage of studded tires was not estimated with the same certainty as at Hornsgatan and Västra Esplanaden, which also significantly affected the quality of the SIMAIR simulation. To summarize, the SVR model greatly improved results at a site difficult for SIMAIR to model.

The SVR for Västra Esplanaden data underestimated the concentration somewhat, compared to the raw SIMAIR model which overestimated. One probable reason for the underestimation was the increase in the use of studded tires, from 88 % and 83 % in 2007 and 2008 to 94 % in 2009. This information was not explicitly included in the SVR; however the SIMAIR model was affected by the percentage and was used as input in the SVR. When the conditions change without chance to learn from the new situation the SVR cannot perform optimally. Several years of data with different use of studded tires might help if the percentage is added explicitly, but those kinds of data were not available for this thesis.

It is not easy to evaluate the uncertainties in the SIMAIR model, but it is important to use the best data available in order to minimize the uncertainty. For example the fact that SIMAIR uses MESAN data every third hour and interpolates it to hourly values is not optimal. This procedure is a remainder of when the quality of the three hour values was better than the quality of the one hour values, before 2006. Now there are high quality hourly values from MESAN available.

When studying the sensitivity of using MESAN data compared to measurements at Torkel Knutssonsgatan in the SVR for Hornsgatan data, the conclusion was that the method was insensitive to if the meteorological data were gridded or came from measurements. A minor improvement was seen when using measured wind direction. The MESAN data were generally good enough at capturing important weather conditions. One needs to keep in mind that there are uncertainties in meteorological observations as well, primarily due to local effects. Wind conditions, for example, can differ a lot between a street surrounded by high buildings and a roof site a few blocks away.

There are uncertainties in the concentration measurements, though the three sites in this thesis have equipment for hourly high quality measurements, located to function well as validation of SIMAIR.

The emission databases are updated every year, but of course there are uncertainties here too. Some local emissions in SIMAIR are estimated using traffic data, and it is difficult to examine uncertainty in traffic data, since often there are no measurements during entire years. The traffic is modelled based on several input parameters, for example yearly daily mean traffic, proportion of heavy traffic and use of studded tires, so high quality traffic measurements improve the traffic model a lot. SLB has very detailed traffic data for

Hornsgatan, so using these time series instead of a general model with some specific inputs will probably give better results. However it is easier to validate the performance of SIMAIR when using a consistent methodology.

A future development is to examine if it is possible to use SVR on sites with few measurements by training on adjacent places. This application is very interesting since air quality models aim at partially complementing measurements.

Parts of this thesis will be submitted to a peer-reviewed scientific journal. The article is written together with the researchers Stefan Andersson and Gunnar Omstedt.

In the coming year the method will be tested for current simulations and hopefully the SVR will then be used on a regular basis.

The results of this Master's Thesis have fulfilled the aims well. A summary of different post-processing methods were discussed. A method using SVR was implemented and analysed for Hornsgatan in Stockholm and validated using data from Västra Esplanaden in Umeå and Gårda in Gothenburg. The SVR performed well for all three sites. The sensitivity in using measured versus gridded meteorological data as input in the SVR was examined and the uncertainties in other components of the air pollution model were discussed.

In conclusion, the results look promising when using SVR as a statistical post-processing method in air pollution modelling. It is important to remember that when conditions change without chance to learn from the new situation the SVR will not perform optimally.

9 References

Andersson, S. and Omstedt, G. (2009). *Validering av SIMAIR mot mätningar av PM10, NO2 och bensen*. Norrköping: SMHI.

Andersson, S. and Omstedt, G. (2012). *Utvärdering av SIMAIR mot mätningar av PM10 och NO2 i Göteborg, Stockholm och Umeå för åren 2006-2009*. Norrköping: SMHI.

Blom, G. et al. (2005). *Sannolikhetsteori och statistikteori med tillämpningar*. Lund: Studentlitteratur.

Burman, L. and Johansson, C. (2010). *Utsläpp och halter av kväveoxider och kvävedioxid på Hornsgatan – Analys av trafikmätningar på Hornsgatan under hösten 2009*. Stockholm: SLB analys vid Miljöförvaltningen i Stockholms stad.

Böiers, L-C. (2010). *Mathematical Methods of Optimization*. Lund: Studentlitteratur AB.

Cristianini, N. and Shawe-Taylor, J. (2012). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Denby, B. and Spangl, W. (2011). *The combined use of models and monitoring for applications related to the European Air Quality Directive: A working sub-group of FAIRMODE*. Norway: NILU, Austria: Federal Environment Agency.

EMEP (2013). *EMEP history and structure* (Online). Available: http://www.emep.int/emep_overview.html(2013-09-10).

EU (2008). Europaparlamentets och rådets direktiv 2008/50/EG av den 21 maj 2008 om luftkvalitet och renare luft i Europa. *Europeiska unionens officiella tidning* L 152/1.

European Commission JRC (2013). *DELTA Air Quality benchmarking tool – general description* (Online). Available: http://aqm.jrc.ec.europa.eu/DELTA/generaldescription.htm (2013-09-12).

European Environment Agency (2011). The application of models under the European Union's Air Quality Directive: A technical reference guide. *EEA Technical report*, 10/2011.

Forum for Air Quality Modelling (2013). *Scope* (Online). Available: http://fairmode.ew.eea.europa.eu/fol568175/scope (2013-09-11).

Gidhagen, L. et al. (2013). High-resolution modeling of residential outdoor particulate levels in Sweden. *Journal of Exposure Science and Environmental Epidemiology* 23, 306-314.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer Science + Business Media.

Häggmark, L. et al. (2000). MESAN, an operational mesoscale analysis system. *Tellus* 52A, 2-20.

Landelius, T., Josefsson, W. and Persson, T. (2001). *A system for modelling solar radiation parameters with mesoscale spatial resolution*. Norrköping: SMHI.

Langner, J. et al. (1998). Validation of the operational emergency response model at the Swedish meteorological and hydrological institute using data from ETEX and the Chernobyl accident. *Atmospheric Environment*, 32(24), 4325-4333.

Lin, H. T., and Lin, C. J. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Submitted to Neural Computation*, 1-32.

Marsland, S. (2009). *Machine Learning – An Algorithmic Perspective*. Palmerston North, New Zealand: CRC Press.

Naturvårdsverket (2012). *Miljömål – Frisk luft – Partiklar (PM10)* (Online). Available: http://www.miljomal.se/sv/Miljomalen/2-Frisk-luft/Preciseringar-av-Frisk-luft/Partiklar-PM10/ (2013-09-11).

Naturvårdsverket (2013a). *Miljömål – Frisk luft – Partiklar i luft* (Online). Available: http://www.miljomal.se/Miljomalen/Alla-indikatorer/Indikatorsida/?iid=105&pl=1(2013-09-10).

Naturvårdsverket (2013b). *Miljökvalitetsnormer* (Online). Available: http://www.naturvardsverket.se/Stod-i-miljoarbetet/Vagledning-amnesvis/Miljokvalitetsnormer/# (2014-01-20).

Omstedt, G. et al. (2012). *Luftkvaliteten i Sverige år 2020*. Norrköping: SMHI.

Osherovich, E. (2010). *Graphic illustrating a convex function* (Online). Available: http://en.wikipedia.org/wiki/File:Epigraph_convex.svg (2014-01-09).

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.

Raabe, O. (1976). Aerosol Aerodynamic Size Conventions For Inertia! Sampler Calibration. *Journal of the Air Pollution Control Association*, 26:9, 856-860.

Refaeilzadeh, P., Tang, L. and Liu, H. (2008). *Cross-Validation*. Arizona State University, USA.

Robertson, L., Langner, J. and Engardt, M. (1999). An Eulerian limited-area atmospheric transport model. *Journal of Applied Meteorology*, 38(2), 190-210.

Scikit learn (2013a). *About us* (Online). Available: http://scikit-learn.org/stable/about.html (2013-11-29).

Scikit learn (2013b). *Support Vector Machines* (Online). Available: http://scikit-learn.org/stable/modules/svm.html(2013-11-29).

Seinfeld, J. and Pandis, S. (2006). *Atmospheric chemistry and physics – From Air Pollution to Climate Change*. Hoboken, New Jersey: John Wiley & Sons.

SMED (2013). *Programområde luft* (Online). Available: http://www.smed.se/luft(2013-09-10).

SMHI (2008). *Data assimilering* (Online). Available: http://www.smhi.se/forskning/forskningsomraden/analys-prognos/dataassimilering-1.176 (2013-09-04).

SMHI (2012). *Detaljerade luftkvalitetskartor visar luftens tillstånd i Umeå* (Online). Available: http://www.smhi.se/2.1209/detaljerade-luftkvalitetskartor-visar-luftens-tillstand-i-umea-1.25271 (2013-12-19).

Smola, A. and Schölkopf, B. (2002). *Learning with Kernels*. Cambridge, Massachusetts, London: The MIT Press.

Smola, A. and Schölkopf, B. (2004). *A Tutorial on Support Vector Regression*. Statistics and computing, 14(3), 199-222.

Svensk författningssamling (2010). *Luftkvalitetsförordning*. SFS 2010:477.

Swedish Environmental Protection Agency (2011). *Sweden's environmental objectives – an introduction*. Stockholm: Swedish Environmental Protection Agency.

Tham, M. T. (2009). *Dealing with measurement noise* (Online). Available: http://lorien.ncl.ac.uk/ming/filter/filewma.htm (2014-01-10).

Thunis, P., Pederzoli, A. and Pernigotti, D. (2012). Performance criteria to evaluate air quality modelling applications. *Atmospheric Environment* 59, 476-482.

Unden, P. et al. (2002). *HIRLAM-5 Scientific documentation*. Norrköping: SMHI.

Vapnik, V. (2010). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Wikipedia (2008). *Graphic showing the maximum separating hyperplane and the margin* (Online). Available: http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png (2013-11-14).

Appendix A

Some optimization theory is given in this appendix. Important definitions and propositions are introduced as a help to understand the theory behind Support Vector Machines, including how to solve a convex optimization problem using duality.

Optimization theory

The definition of a convex function is seen in Definition 1 (Christianini and Shawe-Taylor 2012, p. 81).

Definition 1.

A function f *is called convex if for all* $w, u \in \mathbb{R}^n$ *and for any* $\theta \in (0,1)$,

$$f(\theta w + (1-\theta)u) \leq \theta f(w) + (1-\theta) f(u).$$

If there is a strict inequality in the expression above, f is called strictly convex.

A convex set is defined in Definition 2, and can be interpreted geometrically as for any two points w_1, w_2 belonging to the set Ω, the line segment between them also belongs to the set (Böiers 2010, p. 119).

Definition 2.

A set Ω *in* $\mathbb{R}^n$ *is said to be convex if*

$$w_1, w_2 \in \Omega \Rightarrow \theta w_1 + (1-\theta) w_2 \in \Omega \text{ for all } 0 < \theta < 1.$$

When f is a convex function defined on a convex set Ω, the optimization problem $\min f(w), w \in \Omega$ is called a convex programming problem.

In Figure 30 (Osherovich 2010) a convex function is exemplified. The region above the graph of a convex function is always a convex set, so a line between any two points in the region always belongs to the set.

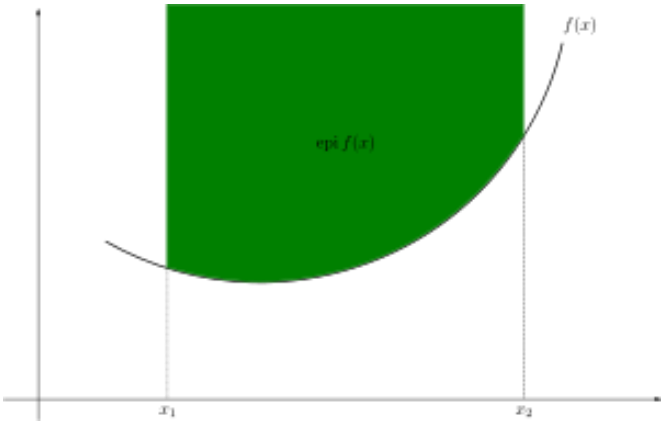

Figure 30. An example of a convex function $f(x)$. The green region above the graph is a convex set.

Optimization with constraints

A general constrained optimization problem, in its primal form, is stated in Definition 3 (Christianini and Shawe-Taylor 2012, p. 80).

Definition 3.

Given functions $f, g_i, i = 1, \dots, k,$ *and* $h_j, j = 1, \dots, m,$ *defined on domain* $\Omega \subseteq \mathbb{R}^n$,

$$minimize\ f(w), w \in \Omega,$$
$$subject\ to\ g_i(w) \leq 0, i = 1, \dots, k,$$
$$h_j(w) = 0, j = 1, \dots, m,$$

where $f(w)$ *is the objective function and* g_i *and* h_j *are inequality and equality constraints. The optimal value of the objective function is called the value of the optimization problem.*

The inequality constraints can be transformed into equality constraints by adding slack variables ξ_i, so $g_i(w) \leq 0 \Leftrightarrow g_i(w) + \xi_i = 0,$ with $\xi_i \geq 0$ (Böiers 2010, p. 159).

If $g_i(w^*) = 0$ for the solution w^*, then the inequality constraint is called active and the slack variable ξ_i is equal to zero (Christianini and Shawe-Taylor 2012, p. 80).

According to Proposition 7 any local optimum that solves the problem $min\, f(w)\,, w \in \Omega, f\ convex$, is a global optimal solution (Böiers 2010, p. 224).

Proposition 7.

Assume that w^* *is a local optimal solution to the problem*

$$\min f(w)\,, w \in \Omega,$$

where $f: \Omega \to \mathbb{R}^n$ *is convex. Then*

1. w^* *is a global optimal solution.*
2. *if* w^* *is a strict local minimum then the global optimal solution is unique.*
3. *if* f *is strictly convex then the global optimal solution* w^* *is unique.*

The following proposition provides a way of dealing with constraints on the optimization problem (Böiers 2010, p. 241).

Proposition 8.

Let f *be a differentiable function defined on an open set* $\Omega \subseteq \mathbb{R}^n$.

Consider the problem of minimizing f *subject to conditions*

$$w \in \Omega,$$
$$g_i(w) \leq 0, i = 1, \dots, k,$$
$$h_j(w) = 0, j = 1, \dots, m,$$

where the constraints are affine:

$$g_i(w) = a_i^T w - c_i, \quad h_j(w) = b_j^T w - d_j.$$

Assume that $\bar{w}$ *is a local minimum point. Then there exist numbers* $\alpha_1, \dots \alpha_k \geq 0,$ *and* $\beta_1, \dots, \beta_m (free)$ *such that*

$$\nabla f(\overline{w}) + \sum_{i=1}^{k} \alpha_i \nabla g_i(\overline{w}) + \sum_{j=1}^{m} \beta_j \nabla h_j(\overline{w}) = 0.$$

For constraints that are not active at $\overline{w}$ the corresponding coefficients α_i are equal to zero.

The proof of this statement uses Farkas' theorem. The parameters α_i and β_j are called Lagrange multipliers. The Lagrange function, also called primal Lagrangian, is defined in Definition 4.

Definition 4.

Given functions $f, g_i, i = 1, \dots, k,$ and $h_j, j = 1, \dots, m,$ defined on domain $\Omega \subseteq \mathbb{R}^n$, consider the problem of minimizing $f(w)$ subject to conditions

$$w \in \Omega,$$
$$g_i(w) \leq 0, i = 1, \dots, k,$$
$$h_j(w) = 0, j = 1, \dots, m.$$

The primal Lagrangian is then defined as

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{m} \beta_j h_j(w), w \in \Omega, \alpha_i \geq 0,$$

where the parameters α_i and β_j are called Lagrange multipliers.

Proposition 8 can also be formulated as:

A necessary condition for $\overline{w}$ to be a local minimum point is that there exist $\alpha \geq 0$ and β $(free)$ such that

$$\nabla_w L(\overline{w}, \alpha, \beta) = 0, \qquad L(\overline{w}, \alpha, \beta) = f(\overline{w}).$$

The results are extended in order to deal with general constraints in the Karush-Kuhn-Tucker (KKT) theorem, here called Proposition 9 (Böiers 2010, p. 250-252).

Proposition 9.

Let f be a differentiable function defined on an open set $\Omega \subseteq \mathbb{R}^n$.
Consider the problem of minimizing f subject to the conditions

$$w \in \Omega,$$
$$g_i(w) \leq 0, i = 1, \dots, k,$$
$$h_j(w) = 0, j = 1, \dots, m.$$

Assume that $\overline{w}$ is a local point of minimum (in particular, that $\overline{w}$ satisfies the constraints). Suppose that the functions g_i are differentiable and the functions h_j are continuously differentiable at $\overline{w}$, and that their gradients at $\overline{w}$ satisfy the following constraint qualification (CQ):

$$(CQ)\begin{cases}\sum_{i=1}^{k}\lambda_i\nabla g_i(\overline{w})+\sum_{j=1}^{m}\mu_j\nabla h_j(\overline{w})=0\\ \lambda_i\geq 0 \text{ for all } i\end{cases}\Rightarrow\begin{cases}\lambda_i=0 \text{ for all } i\\ \mu_j=0 \text{ for all } j.\end{cases}$$

Then there exist scalars α_i and β_j such that

$$\nabla_w L(\overline{w},\alpha,\beta)=0, \quad (1)$$

$$\alpha\geq 0, \quad (2)$$

$$\alpha_i g_i(\overline{w})=0 \; for\; all\; i. \quad (3)$$

It can be noted that the constraint qualifications only need to be fulfilled for constraints active at $\overline{w}$. The conditions implied by $\overline{w}$ are called feasibility conditions, while constraints (1) and sometimes (2) are called KKT conditions and constraints (3) are called complementary slackness conditions and imply that $\alpha_i = 0$ for all non-active inequality constraints $g_i(\overline{w})$.

The points $\overline{w}$ that satisfy feasibility conditions and the constraints (1)-(3) are called KKT points. If there are any points that do not fulfil the constraint qualifications these are called CQ points. If a global minimum is known to exist (for example due to compactness) it is among the KKT or CQ points. So by checking the function values of all these points the minimum is found. Being a KKT point is a necessary condition for a local minimum, but not a sufficient condition (Böiers 2010, p. 252).

Sufficient conditions for a minimum are stated in Proposition 10 (Böiers 2010, p. 264), where $I=\{i;\ g_i(\overline{w})=0\}$.

Proposition 10.

Assume that $\overline{w}\in\Omega$ is a KKT point for the minimum problem stated in Proposition 9.

Also assume that the functions f and $g_i, i\in I$, are convex in a neighbourhood of $\overline{w}$,

and that all h_j are affine there.

Then f has a local minimum at $\overline{w}$.

This result can be further strengthened in order to apply to global minima (Böiers 2010, p. 265).

Proposition 11.

If we strengthen the assumptions in Proposition 10 *to convexity of the set Ω, convexity of f and of all g_i in all of Ω and affinity of h_j in Ω,*

then f has a global minimum at $\overline{w}$.

When the conditions above are fulfilled, finding a KKT point is equivalent to finding a global minimum.

Duality

An introduction to the concept of duality is presented here. First a saddle point is defined in Definition 5 (Böiers 2010, p. 294).

Definition 5.

A point $(\overline{w},\bar{\alpha},\bar{\beta})\in\Omega\times U$, where $U=\{(\alpha,\beta)\in\mathbb{R}^k\times\mathbb{R}^m;\alpha\geq 0\}$, is called a

$saddle\ point\ of\ the\ Lagrangian\ L\ if$

$L(\bar{w}, \alpha, \beta) \leq L(\bar{w}, \bar{\alpha}, \bar{\beta}) \leq L(w, \bar{\alpha}, \bar{\beta}) for\ all\ (w, \alpha, \beta) \in \Omega \times U.$

In Proposition 12 the necessary conditions for a saddle point are stated (Böiers 2010, p. 294).

Proposition 12.

$The\ point\ (\bar{w}, \bar{\alpha}, \bar{\beta}) \in \Omega \times U\ is\ a\ saddle\ point\ of\ L\ if\ and\ only\ if\ the$

$following\ three\ conditions\ are\ true.$

1. $L(\bar{w}, \bar{\alpha}, \bar{\beta}) = min\, L(w, \bar{\alpha}, \bar{\beta}), \qquad w \in \Omega.$

2. $g(\bar{w}) \leq 0\ and\ h(\bar{w}) = 0.$

3. $\alpha_i g_i(\bar{w}) = 0\ for\ all\ i.$

A saddle point is therefore a feasible point to the optimization problem stated in Definition 3 and $L(\bar{w}, \bar{\alpha}, \bar{\beta}) = f(\bar{w})$. The following interesting conclusions are presented in Proposition 13 (Böiers 2010, p. 296).

Proposition 13.

$Assume\ that\ (\bar{w}, \bar{\alpha}, \bar{\beta})\ is\ a\ saddle\ point\ of\ L.\ Then$

1. $\bar{w}\ solves\ the\ minimization\ problem.$

2. $\bar{w}, \bar{\alpha}, \bar{\beta}\ satisfy\ the\ KKT\ conditions.$

The parameters $(\bar{\alpha}, \bar{\beta})$ that occur in the definition of a saddle point are exactly the same as the Lagrange multipliers in Proposition 8.

The following proposition states sufficient conditions for a saddle point.

Proposition 14.

$Assume\ that$

1. $the\ set\ \Omega\ is\ convex,$

2. $the\ functions\ f\ and\ g_i, i = 1, .., k\ are\ convex\ on\ \Omega,$

3. $the\ functions\ h_j, j = 1, \dots, m\ are\ affine\ on\ \Omega,$

4. $\bar{w}, \bar{\alpha}, \bar{\beta}\ satisfy\ the\ KKT\ conditions, and\ \bar{w}\ is\ feasible\ for\ the\ primal\ problem.$

$Then\ (\bar{w}, \bar{\alpha}, \bar{\beta})\ is\ a\ saddle\ point\ of\ the\ Lagrange\ function\ L.$

Now the Lagrangian dual problem to the primal optimization problem stated in Definition 3 is defined below (Böiers 2010, p. 297). Note that the objective function $W(\alpha, \beta)$ contains the equality and inequality constraints from the primal problem.

Definition 6.

$The\ Lagrangian\ dual\ problem\ to\ the\ problem\ in$ Definition 3 is

$$maximize\ W(\alpha, \beta) = \inf_{w \in \Omega} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{m} \beta_j h_j(w) = \inf_{w \in \Omega} L(w, \alpha, \beta)$$

$$subject\ to\ (\alpha, \beta) \in U, \qquad U = \{(\alpha, \beta) \in \mathbb{R}^k \times \mathbb{R}^m; \alpha \geq 0\}.$$

The function inf is defined as the largest real number that is smaller than or equal to every number in the set. The relation $W(\alpha, \beta) \leq L(w, \alpha, \beta) \leq f(w)$ holds according to Definition 3 and Definition 6. So if there are any feasible points the relation

$$\sup_{(\alpha,\beta) \in \mathrm{U}} W(\alpha, \beta) \leq \inf_{w \in \Omega} f(w)$$

holds, and the difference between the two sides are called the duality gap. If the duality gap is equal to zero there is strong duality. If there are $\overline{w}$ and $(\bar{\alpha}, \bar{\beta})$, all feasible for both the primal and the dual problem, then $f(\overline{w}) = W(\bar{\alpha}, \bar{\beta}) = L(\overline{w}, \bar{\alpha}, \bar{\beta})$, as formulated in Proposition 15.

Proposition 15.

The point $(\overline{w}, \bar{\alpha}, \bar{\beta}) \in \Omega \times U$ *is a saddle point of* L *if and only if* $\overline{w}$ *is feasible for the primal problem and*

$$f(\overline{w}) = W(\bar{\alpha}, \bar{\beta}).$$

A strong duality that assures that the primal and the dual problem have the same value is stated in Proposition 16 (Christianini and Shawe-Taylor 2012, p. 86).

Proposition 16.

Given an optimization problem on a convex set $\Omega \subseteq \mathbb{R}^n$,

$$minimize\ f(w), w \in \Omega,$$
$$subject\ to\ g_i(w) \leq 0, i = 1, \dots, k,$$
$$h_j(w) = 0, j = 1, \dots, m,$$

where g_i *and* h_j *are affine functions:* $g_i(w) = a_i^T w - c_i, \qquad h_j(w) = b_j^T w - d_j,$

the duality gap is zero.

In conclusion, when having a convex optimization problem strong duality assures that solving the dual problem is equivalent to solving the primal problem. This is very useful in cases when the dual problem is more easily solved.

Appendix B

In Tables B1, B2 and B3 the correlations between model error and different covariates from Hornsgatan data in 2007 and 2008, using Spearman's rank, are provided. For example the exponentially filtered precipitation during winter has a lot higher correlation to the model error using Spearman's rank compared to the correlation computed by Pearson's r.

Table B1. The significant correlations between model error and humidity, exponentially filtered precipitation and GI from Hornsgatan in 2007 and 2008, using Spearman's rank.

	Humidity obs.	Humidity MESAN	Precipitation obs. (exp. filtered)	Precip-itation MESAN (exp. filtered)	GI obs.	GI STRÅNG
2007 - 2008 winter period:	-0.31	-0.37	0.34	0.28	0.03	-
2007 - 2008 spring period:	-0.22	-0.23	0.05	0.04	0.12	0.09
2007 - 2008 summer period:	0.09	0.09	0.13	0.21	-	-

Table B2. The significant correlations between model error and concentration in urban background, total number of vehicles, heavy and light vehicles and total emission from Hornsgatan in 2007 and 2008, using Spearman's rank.

	PM10 urban background obs.	PM10 urban background mod.	Nbr of vehicles	Nbr of heavy vehicles	Nbr of light vehicles	Total emission
2007 - 2008 winter period:	0.28	-0.32	0.05	-	0.05	-0.15
2007 - 2008 spring period:	0.51	-0.08	0.16	0.17	0.16	0.18
2007 - 2008 summer period:	0.36	-0.07	0.04	0.14	-	-0.10

Table B3. The significant correlations between model error and temperature, u- and v-components of wind and wind direction from Hornsgatan in 2007 and 2008, using Spearman's rank.

	Temp. obs.	Temp. MESAN	u-wind obs.	u-wind MESAN	v-wind obs.	v-wind MESAN	Wind dir. obs.	Wind dir. MESAN
2007 - 2008 winter period:	0.12	0.13	-0.10	-0.08	0.06	0.05	-0.04	-0.04
2007 - 2008 spring period:	0.13	0.13	-0.03	-0.07	0.16	0.16	0.13	0.10
2007 - 2008 summer period:	-0.09	-0.10	-0.13	-0.18	0.12	0.10	0.08	0.03

Appendix C

The resulting C and γ found by validation for each of the three regressions, for all sites are seen in Table C1. Each pair of parameters is used to find the optimal hyperplane, which is later used for prediction. It can be noted that the parameters do not differ that much (remember that they are found using a log-scale) between the sites, which may indicate that the method generalizes well.

Table C1. The resulting C and γ found by validation for each of the three regressions. The first column shows results for Hornsgatan data with MESAN data. The second column shows results for Hornsgatan with measurements of wind direction instead of MESAN data. The third column presents results for Västra Esplanaden and the forth column for Gårda. Data are from 2007 and 2008.

	Hornsgatan		Hornsgatan (met. obs.)		V. Esplanaden		Gårda	
	C	γ	C	Γ	C	γ	C	γ
Low interval (0-25 µg/m^3)	128	2	128	2	32	2	32	2
Medium (25-50 µg/m^3)	128	2	128	2	32	2	32	2
High (>50 µg/m^3)	128	1/8	128	½	128	1/2	128	1/2

Appendix D

Results from the Support Vector Regression, using observations of wind direction instead of MESAN data, are gathered in Table D1. The improvement is a slightly lower RMSE compared to using MESAN wind direction.

Table D1. A summary of the results of the SVR compared to the SIMAIR model at Hornsgatan in 2009, with observed wind direction as input parameter instead of MESAN wind direction.

	Observations	SIMAIR model	SVR
Yearly mean (µg/m^3)	37.2	28.2	33.8
90-percentile daily mean (µg/m^3)	81.9	59.3	66.5
Days > 50 µg/m^3	67	46	54
RPE %		24	9.1
RDE %		22	8.4
r daily mean		0.73	0.79
r hourly		0.63	0.68
RMSE daily mean (µg/m^3)		14.2	12.4
RSME hourly (µg/m^3)		36.0	32.9

Master's Theses in Mathematical Sciences 2014:E2
ISSN 1404-6342

LUTFMS-3235-2014

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/